# Clustering de trajectoires de traitement

*Application à une population de patients diabétiques*

Stage de Romane Le Goff - Mars à Août 2022
Encadré par Geoffray Bizouard et Oriane Bretin (IQVIA)

# Sujet du stage (1/2)

**1 Contexte**

Le **parcours de soins** : **prise en charge** du patient dans le temps
(exemple : consultation généraliste, puis prescription, puis hospitalisation...)

Intérêt : **regrouper** les patients aux **parcours de soins similaires**
(décrire, catégoriser les parcours, repérer des profils spécifiques à cibler)

**2 Objectifs**

**Evaluer** des **méthodes de clustering** permettant de **regrouper les patients** aux trajectoires de **traitement similaires**

**3 Base de données**

**Délivrances de traitements** dans un panel de 9600 **pharmacies** en France – 40 millions de patients

**4 Population d'étude**

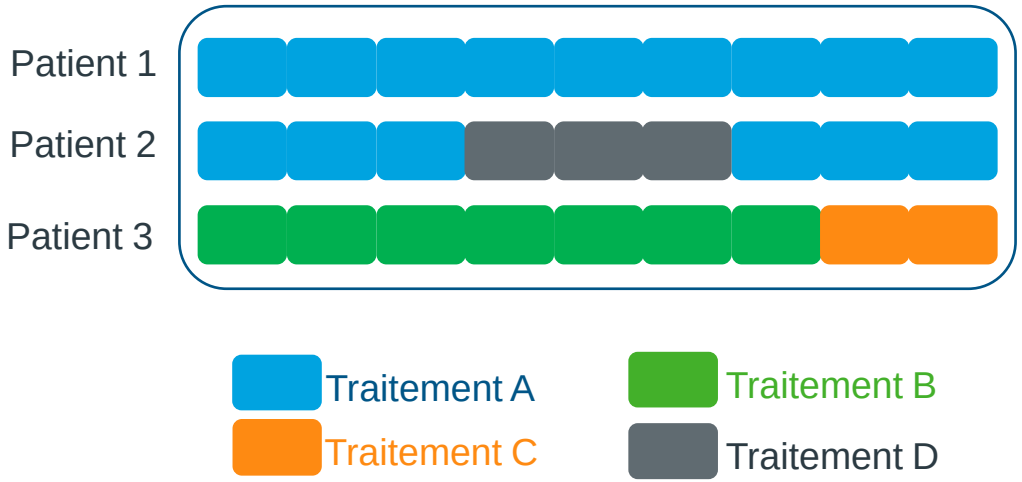
Population de **patients diabétiques**

# Sujet du stage (2/2)

*Notions clés*

## 1. Distances entre patients

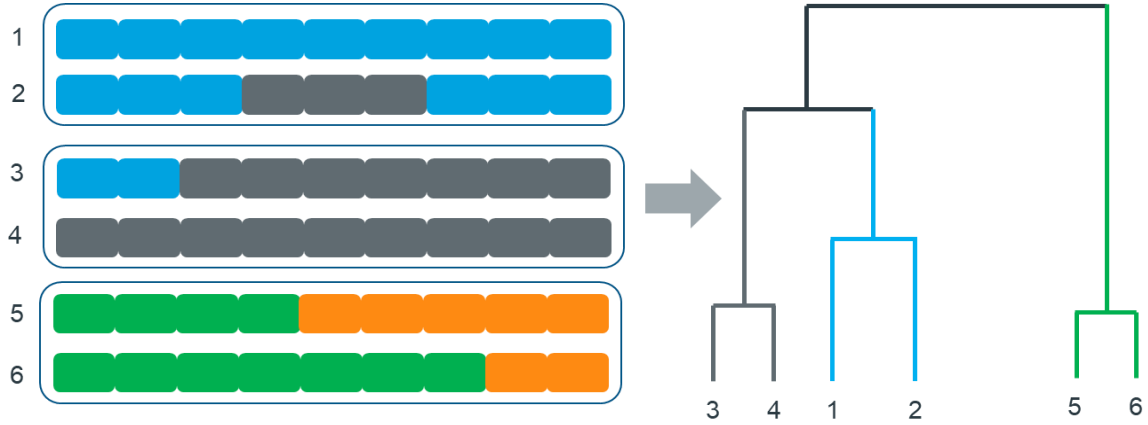
- Que signifie « trajectoires **similaires** » ?
- Comment définir des **distances entre patients** ?
- Les patients 1 et 2 semblent proches mais comment le mesurer ?




**→ Le calcul de distances est le point central du stage**

## 2. Clustering

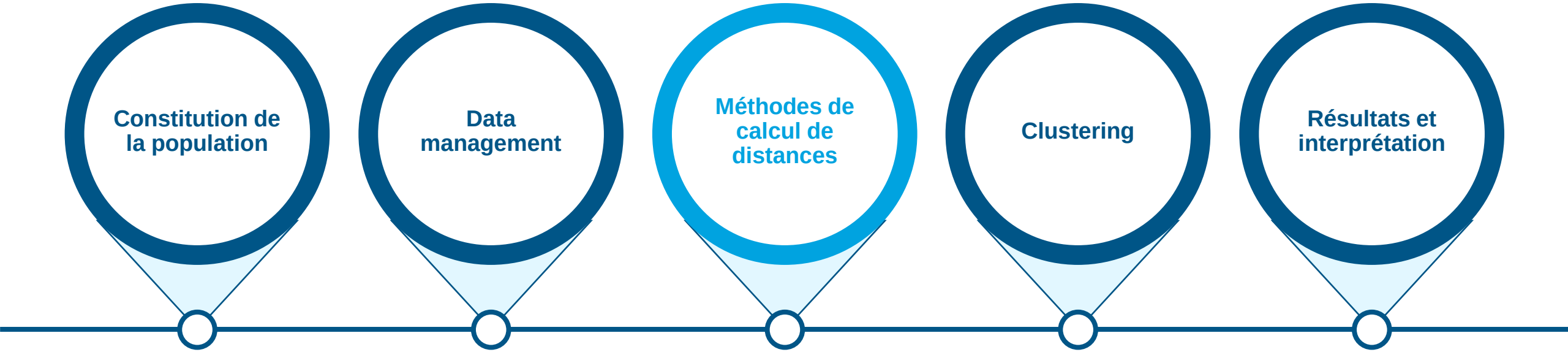
**Méthodes de clustering**
Classification Ascendante Hiérarchique ou autre méthode




**Choix des groupes**
Coupure du dendogramme

# Aperçu du stage

*En 5 étapes*




Population de patients diabétiques

Extraction et reconstitution des **enchainements de traitements**

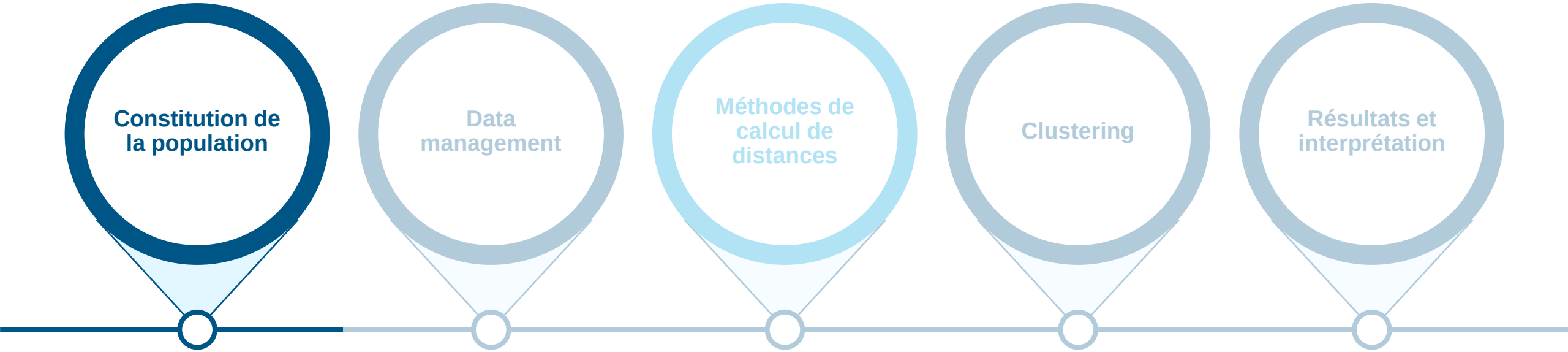
**Optimal Matching**
(d'autres méthodes ont été explorées)

Classification Ascendante hiérarchique (CAH)
Partitioning Around Medoids (PAM)

Description des groupes

# Aperçu du stage

*En 5 étapes*




Population de patients diabétiques

Extraction et reconstitution des **enchainements de traitements**

**Optimal Matching**
(d'autres méthodes ont été explorées)

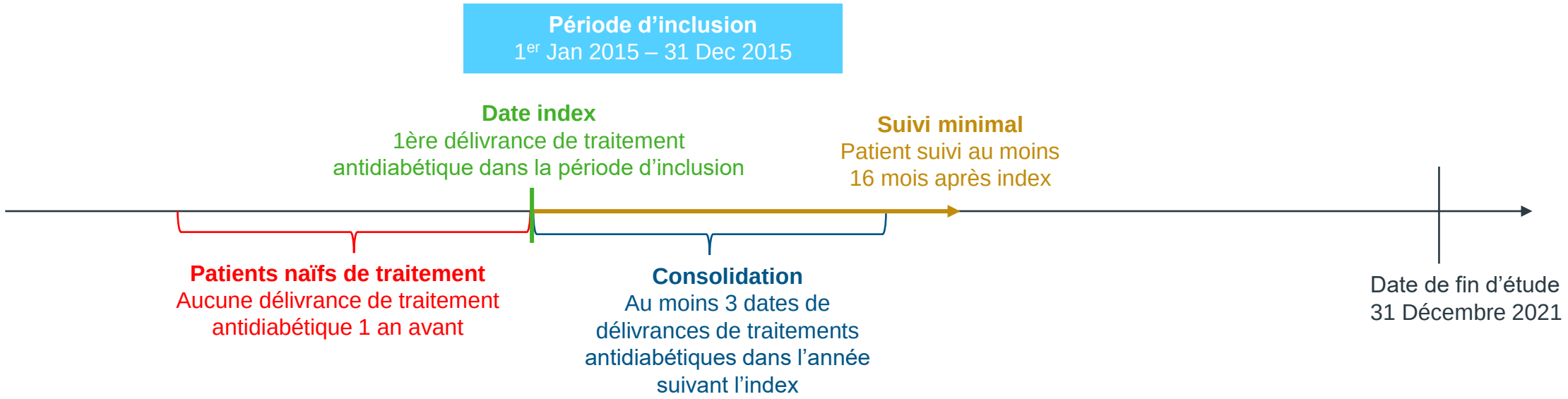
Classification Ascendante hiérarchique (CAH)
Partitioning Around Medoids (PAM)

Description des groupes

# Constitution de la population

*Patients diabétiques naïfs de traitement*

## Critères de sélection principaux



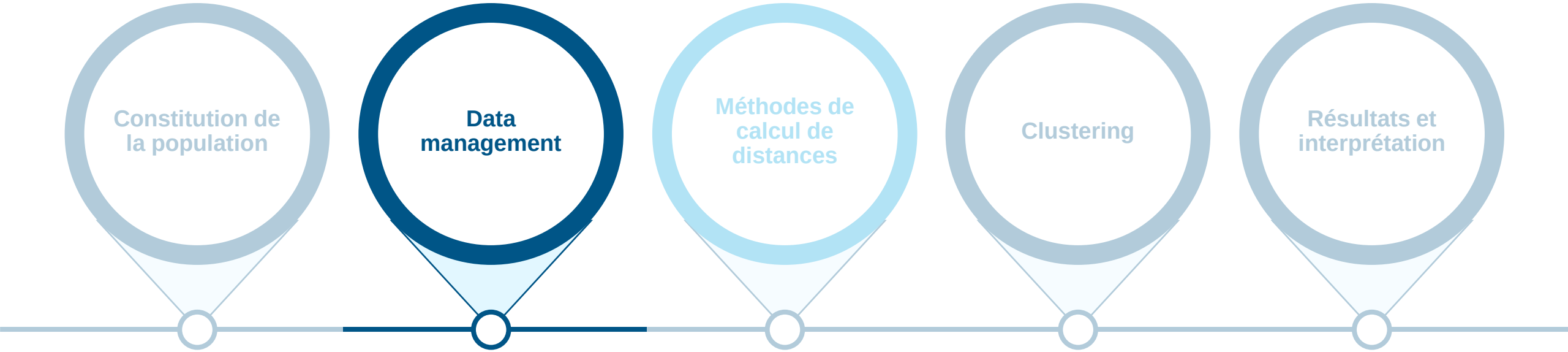

## Critères additionnels

- Critères pour s'assurer que le patient est bien **suivi** régulièrement par une pharmacie du panel, avant et après inclusion

**Total N = 16 843**

# Aperçu du stage

*En 5 étapes*




Population de patients diabétiques

Extraction et reconstitution des **enchainements de traitements**

**Optimal Matching**
(d'autres méthodes ont été explorées)

Classification Ascendante hiérarchique (CAH)
Partitioning Around Medoids (PAM)

Description des groupes

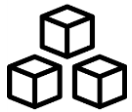
# Data management (1/2)

*Reconstruction de l'enchaînement des traitements*

## En entrée dans la base…

Produit délivré

Nombre de boîtes



Nombre de comprimés par boites

## Data management

**Estimation de la durée du traitement**

- A partir de la Defined Daily Dose (DDD) (posologie moyenne)
- Pour l'insuline, 1 stylo = 30 jours

**Choix des états**

Monothérapie, Bithérapie, Trithérapie, Insuline, Insuline + OAD, Sans rien

**Construction des trajectoires**

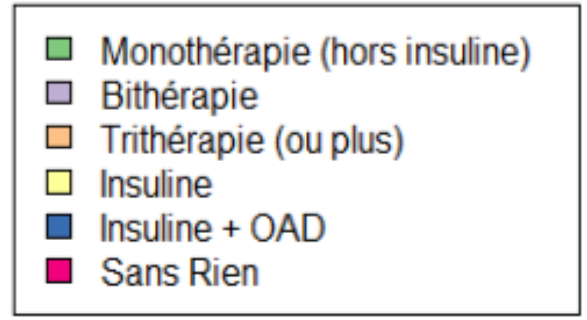
Traitement majoritaire sur des <u>périodes de 4 mois</u>

## Sortie

| ID Patient | Période 1 | Période 2 | … | Période 17 | Période 18 |
|---|---|---|---|---|---|
| 1 | Monothérapie | Monothérapie | … | Bithérapie | Bithérapie |
| 2 | Insuline | Insuline | … | Insuline | Insuline |
| 3 | Monothérapie | Bithérapie | …. | Trithérapie | **NA (fin de suivi)** |
| 4 | Monothérapie | **Sans Rien** | | **Sans Rien** | **Sans Rien** |

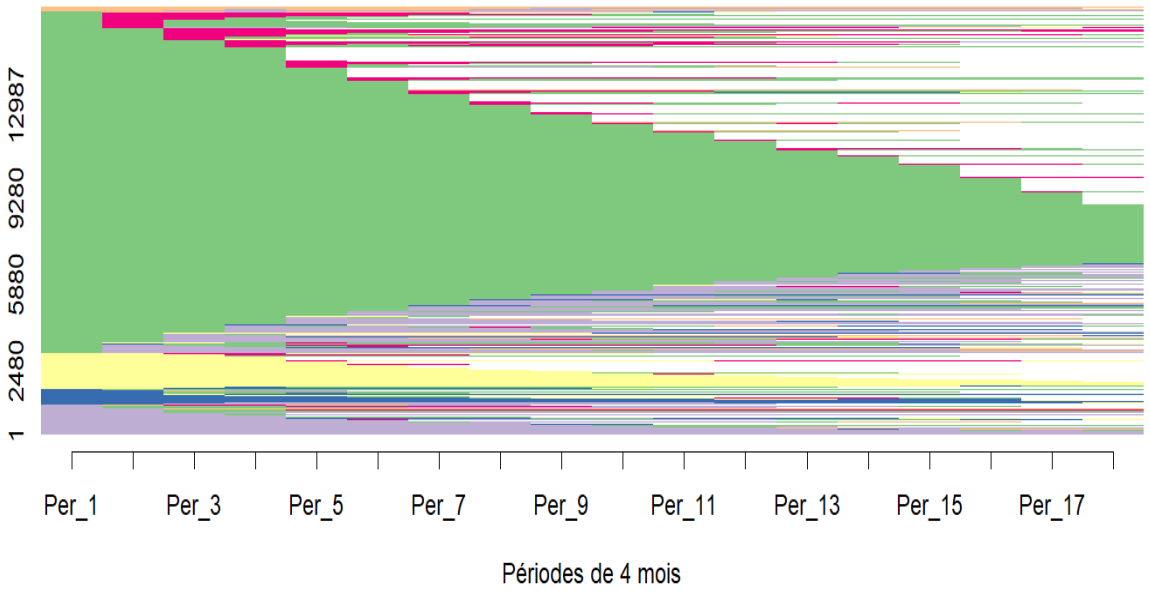
# Data management (2/2)

*Illustration des trajectoires à l'issue du data management*



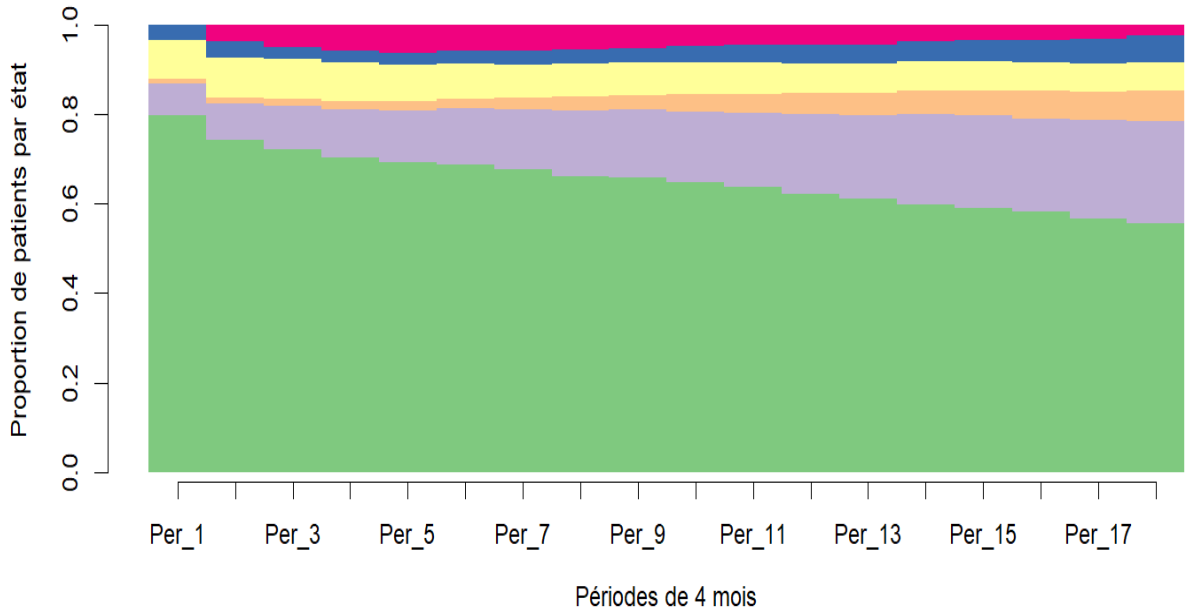

**Tapis des trajectoires de traitement**

Une ligne = un patient




**Chronogramme**

A chaque période, proportion de patients dans chaque état

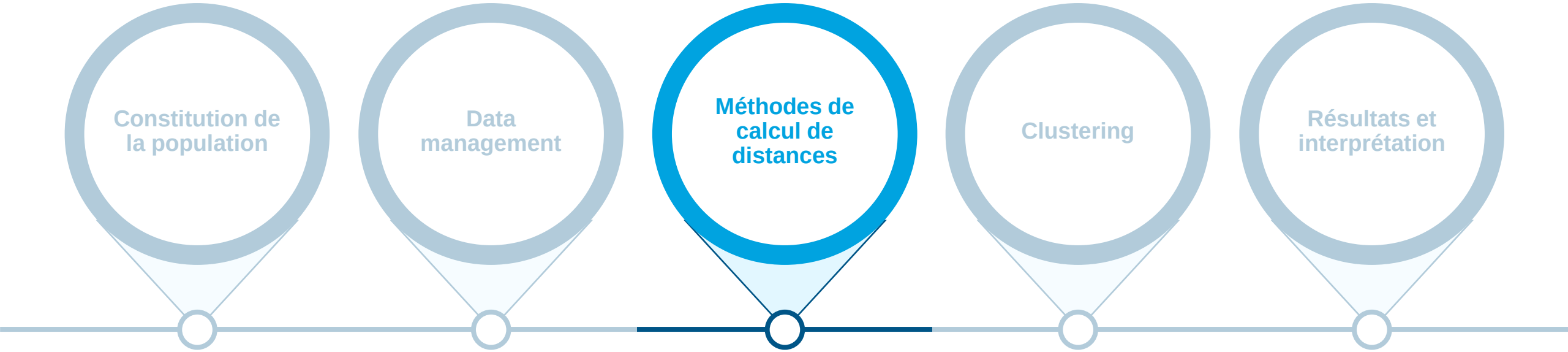



**Remarque**

- On différencie une **fin de suivi** (blanc), d'une **absence de traitement** en étant toujours suivi (modalité Sans Rien)

# Aperçu du stage

*En 5 étapes*




Population de patients diabétiques

Extraction et reconstitution des **enchainements de traitements**

**Optimal Matching**
(d'autres méthodes ont été explorées)

Classification Ascendante hiérarchique (CAH)
Partitioning Around Medoids (PAM)

Description des groupes

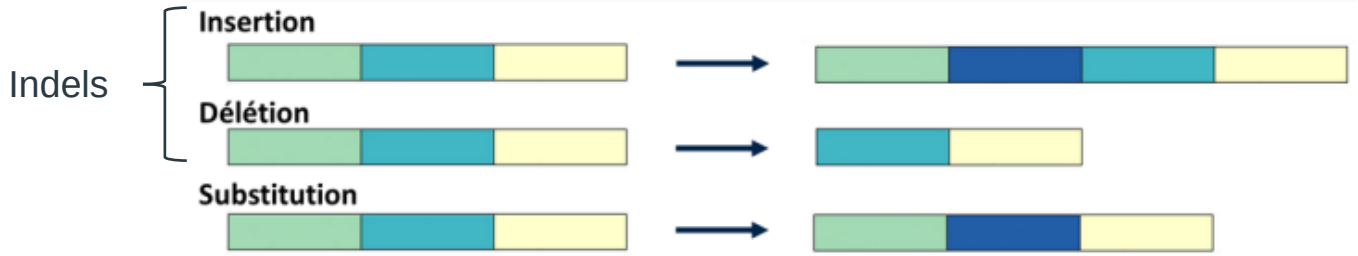
# Méthode de calcul de distances

*Optimal Matching (méthode principale)*

### Principe de la méthode

- On définit des **opérations élémentaires,** inspirées de la Bio-informatique (mutations de l'ADN) et on leur attribue chacune un **coût**

Opérations élémentaires




Opérer, ça coûte !

- **Distances entre 2 patients = somme des coûts des opérations minimales** nécessaires pour rendre les 2 trajectoires identiques

Comment rendre ces 2 trajectoires identiques ?

# Méthode de calcul de distances

*Optimal Matching (méthode principale)*

**Illustration**

Comment passer du patient 1 au patient 2 ?

Patient 1

Patient 2

**Plusieurs possibilités !**

1

Patient 1

Patient 2

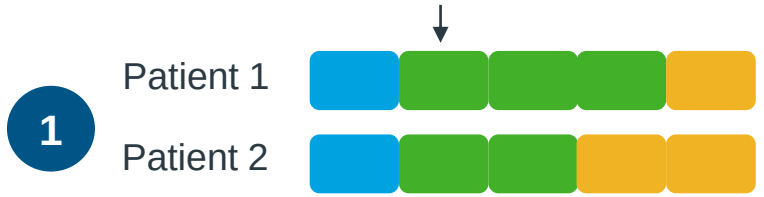
# Méthode de calcul de distances

*Optimal Matching (méthode principale)*

**Illustration**

Comment passer du patient 1 au patient 2 ?


**Plusieurs possibilités !**




Une substitution

# Méthode de calcul de distances

*Optimal Matching (méthode principale)*

**Illustration**

Comment passer du patient 1 au patient 2 ?


**Plusieurs possibilités !**

Et une deuxième substitution

1

# Méthode de calcul de distances

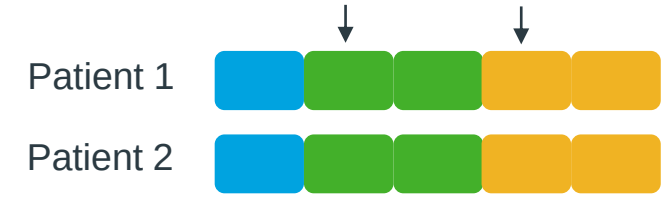
*Optimal Matching (méthode principale)*

**Illustration**

Comment passer du patient 1 au patient 2 ?


**Plusieurs possibilités !**

1




Ou bien …

2

# Méthode de calcul de distances

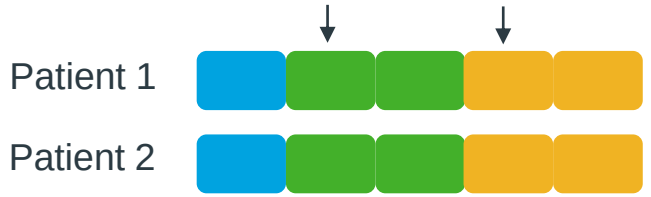
*Optimal Matching (méthode principale)*

**Illustration**

Comment passer du patient 1 au patient 2 ?

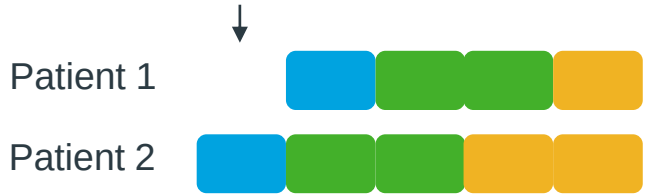

**Plusieurs possibilités !**

1




2

Une délétion



# Méthode de calcul de distances

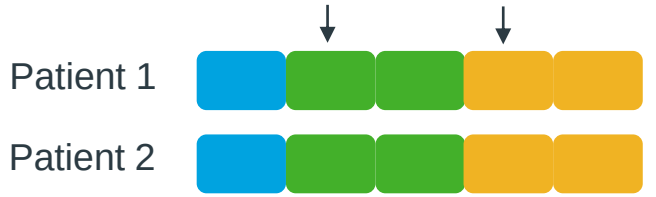
*Optimal Matching (méthode principale)*

**Illustration**

Comment passer du patient 1 au patient 2 ?


**Plusieurs possibilités !**

1




2

Puis une insertion

# Méthode de calcul de distances

*Optimal Matching (méthode principale)*

**Illustration**

Quelle transformation choisir parmi les 2 ?

→ Tout dépend du **système de coût fixé**
→ On s'oriente vers le **coût minimal**

Exemple d'un système de coût

| Indels | Substitution |
|---|---|
| 0.5 | 1 |

**1** 2 substitutions → **coût = 2**

**2** 1 délétion + 1 insertion → **coût = 1**

→ Le choix du **système de coût** peut fortement **impacter les distances** entre individus et donc le **clustering**
  - Coûts attribués **manuellement**, ou
  - Coûts construits **automatiquement** à partir des données (plusieurs stratégies sur R)

→ L'Optimal Matching a l'avantage d'autoriser des **distorsions temporelles** via les opérations d'indels.

# Méthode de calcul de distances

*Optimal Matching - Comparaison de 2 méthodes de calcul de coûts*

**1. Coûts de substitution basés sur les taux de transition entre états** (coût d'indels unique = *max(coûts substitution)/2*)

| | Monothérapie | Bithérapie | Insuline | Insuline + OAD | Trithérapie | Sans Rien |
|---|---|---|---|---|---|---|
| Monothérapie | 0 | 1.909 | 1.992 | 1.950 | 1.970 | 1.712 |
| Bithérapie | 1.909 | 0 | 1.996 | 1.965 | 1.893 | 1.970 |
| Insuline | 1.992 | 1.996 | 0 | 1.903 | 1.998 | 1.951 |
| Insuline + OAD | 1.950 | 1.965 | 1.903 | 0 | 1.969 | 1.994 |
| Trithérapie | 1.970 | 1.893 | 1.998 | 1.969 | 0 | 1.992 |
| Sans Rien | 1.712 | 1.970 | 1.951 | 1.994 | 1.992 | 0 |

Si l'état **A** est **souvent suivi** de l'état **B** dans la population (ou réciproquement), alors :
**coût de substitution AB faible**.

**2. Coûts basés sur les indels**

- Coûts d'indels dépendant de la fréquence de l'état : $C_I(i) = \frac{1}{f_i}$ où $f_i$ est la fréquence relative de l'état *i*

| Monothérapie | Bithérapie | Insuline | Insuline + OAD | Trithérapie | Sans Rien |
|---|---|---|---|---|---|
| 1.990 | 9.375 | 17.604 | 36.731 | 40.887 | 29.543 |

Il coûtera **moins cher** d'insérer/supprimer un état **fréquent**

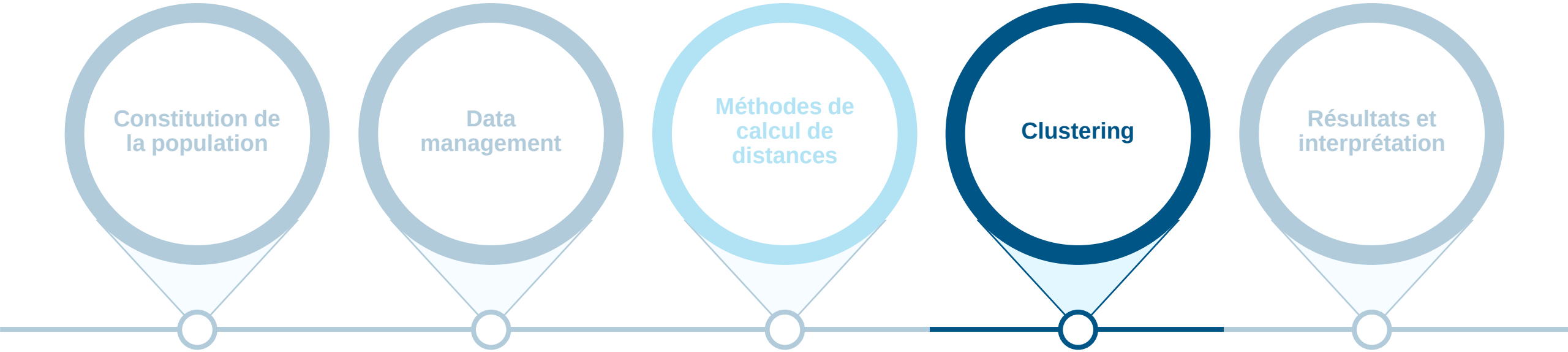
- Coûts de substitution égaux à la somme des coûts d'indels des deux états : $C_S(i,j) = C_S(i) + C_S(j)$

| | Monothérapie | Bithérapie | Insuline | Insuline + OAD | Trithérapie | Sans Rien |
|---|---|---|---|---|---|---|
| Monothérapie | 0 | 11.366 | 19.594 | 38.721 | 42.877 | 31.534 |
| Bithérapie | 11.366 | 0 | 26.979 | 46.106 | 50.262 | 38.919 |
| Insuline | 19.594 | 26.979 | 0 | 54.334 | 58.490 | 47.147 |
| Insuline + OAD | 38.721 | 46.106 | 54.334 | 0 | 77.617 | 66.274 |
| Trithérapie | 42.877 | 50.262 | 58.490 | 77.617 | 0 | 70.430 |
| Sans Rien | 31.534 | 38.919 | 47.147 | 66.274 | 70.430 | 0 |

Il coûtera **moins cher** de substituer 2 états dès lors que ceux-ci sont **fréquents**
Matrices symétriques !

# Aperçu du stage

*En 5 étapes*




Population de patients diabétiques

Extraction et reconstitution des **enchainements de traitements**

**Optimal Matching**
(d'autres méthodes ont été explorées)

Classification Ascendante hiérarchique (CAH)
Partitioning Around Medoids (PAM)

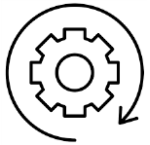
Description des groupes

# Clustering

*En quelques mots*

**Deux méthodes utilisées**

- CAH
- PAM



**Choix du nombre de groupes**

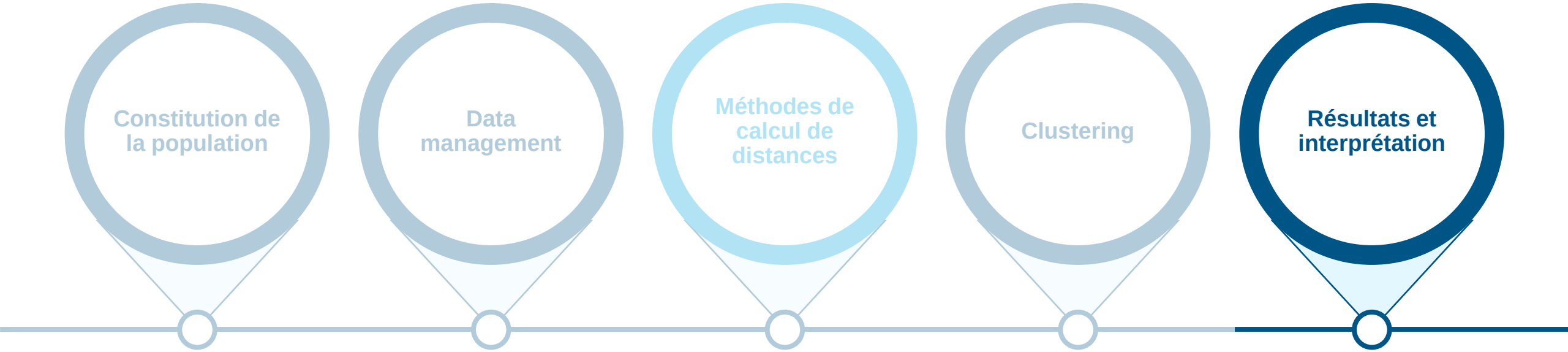
- Le nombre de groupes optimal a été guidé par des **considérations épidémiologiques**
  - Combien de trajectoires différentes sont attendues dans une population diabétique ?

**Validation des groupes**

- Guidée par des **considérations épidémiologiques**
  - Les groupes paraissent-ils homogènes en termes de trajectoires et profils (Age, sexe, traitement majoritaire) ?

# Aperçu du stage

*En 5 étapes*




Population de patients diabétiques

Extraction et reconstitution des **enchainements de traitements**

**Optimal Matching**
(d'autres méthodes ont été explorées)

Classification Ascendante hiérarchique (CAH) Partitioning Around Medoids (PAM)

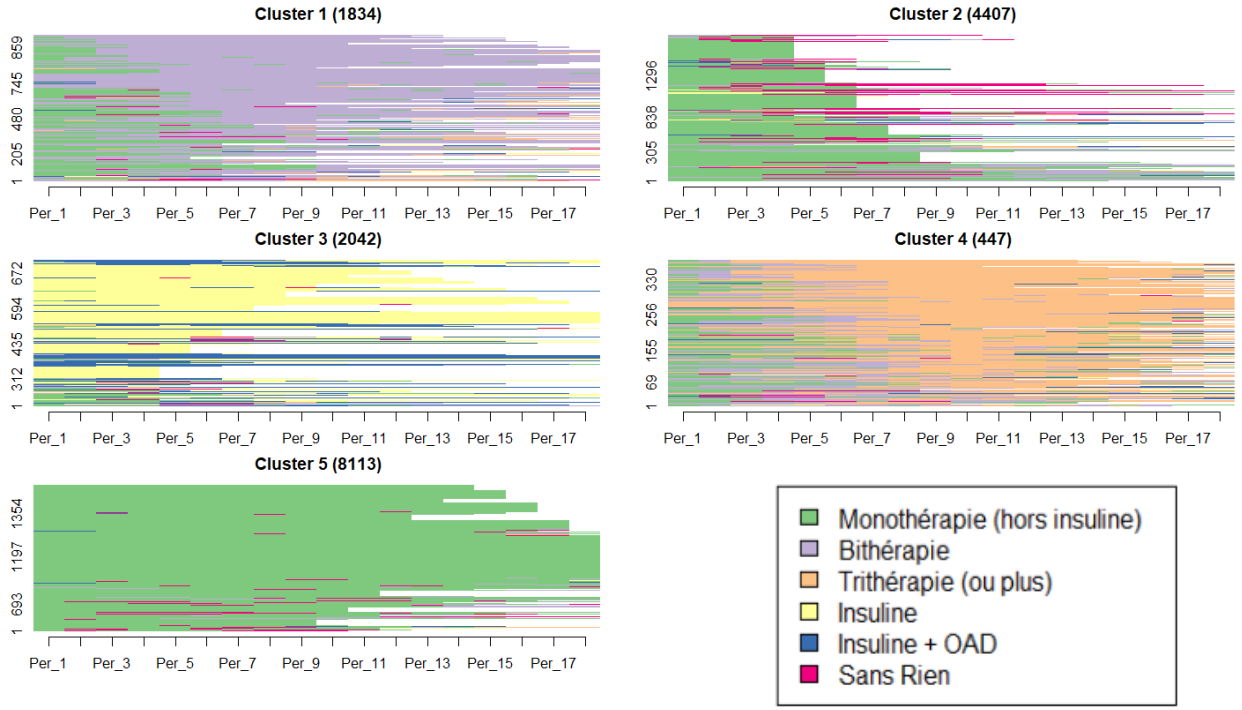
Description des groupes

# Résultats et interprétation

*Optimal Matching - 5 groupes - 2 méthodes de calcul de coûts*

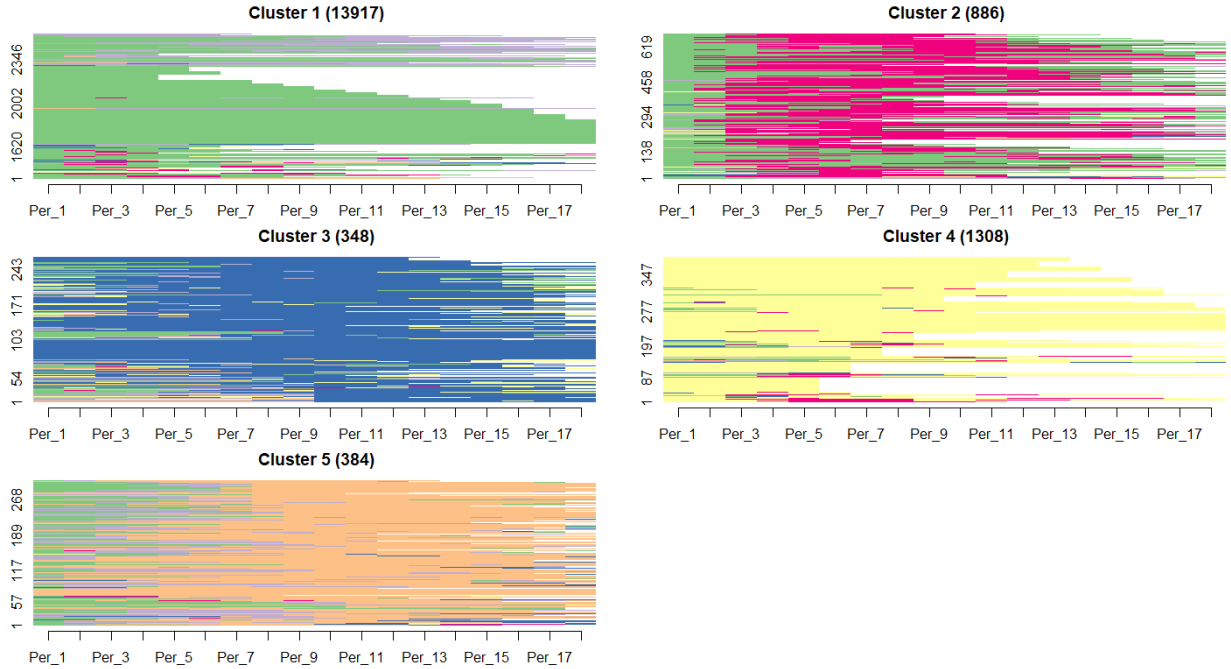
## 1. Coûts basés sur les taux de transition entre états

*Il est peu cher de substituer deux états qui se succèdent souvent dans la population*




## 2. Coûts basés sur les indels

*Les états plus fréquents coûtent moins chers à insérer*




→ **1ère méthode** : Cluster 2 = faible durée de suivi (→ coût d'indel unique)
→ **2ième méthode** : Trithérapies et Insuline+OAD sont chers à insérer et substituer ! Donc ils sont séparés. Plus grand déséquilibre d'effectifs.

# Résultats et interprétation

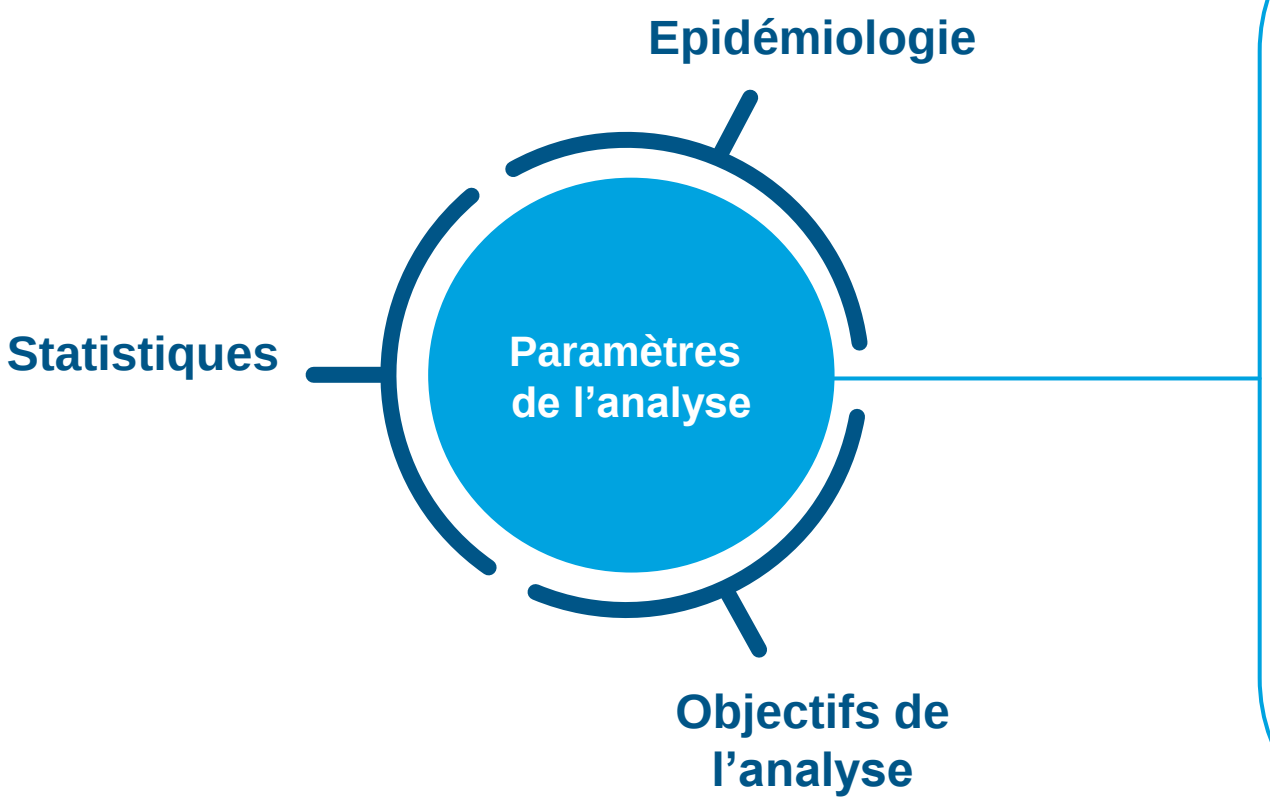
*Décrire les profils patients dans chaque groupe*

| | Cluster 1<br>Mono et bithérapie<br>(N=13917) | Cluster 2<br>Insuline + OAD<br>(N=886) | Cluster 3<br>Sans Rien<br>(N=348) | Cluster 4<br>Insuline<br>(N=1308) | Cluster 5<br>Trithérapie et +<br>(N=384) | Total<br>(N=16843) |
|---|---|---|---|---|---|---|
| Sexe | | | | | | |
| Femme | 5,843 (42%) | 432 (49%) | 140 (40%) | 701 (54%) | 128 (33%) | 7,244 (43%) |
| Homme | 7,849 (56%) | 440 (50%) | 205 (59%) | 573 (44%) | 252 (66%) | 9,319 (55%) |
| Non connu | 225 (1.6%) | 14 (1.6%) | 3 (0.9%) | 34 (2.6%) | 4 (1.0%) | 280 (1.7%) |
| Âge | | | | | | |
| >70 | 4,857 (35%) | 294 (33%) | 96 (28%) | 606 (46%) | 44 (11%) | 5,897 (35%) |
| 55-70 | 6,383 (46%) | 394 (44%) | 137 (39%) | 280 (21%) | 207 (54%) | 7,401 (44%) |
| 35-55 | 2,476 (18%) | 177 (20%) | 111 (32%) | 233 (18%) | 130 (34%) | 3,127 (19%) |
| 18-35 | 184 (1.3%) | 21 (2.4%) | 3 (0.9%) | 125 (9.6%) | 3 (0.8%) | 336 (2.0%) |
| <18 | 17 (0.1%) | 0 (0%) | 1 (0.3%) | 64 (4.9%) | 0 (0%) | 82 (0.5%) |

*Résultats Optimal Matching – Coûts basés sur les indels – 5 groupes*

**Cluster 4** : 2 profils
- Patients âgés (DT2)
- Patient très jeunes (probablement DT1)

# Retour sur les paramètres de l’analyse




## Choix méthodologiques pour l’Optimal Matching

- Choix du **pas de la discrétisation temporelle** (4 mois ici)
- Choix du **nombre d’états** dans la classification (6 modalités de traitement ici)
- Choix des **coûts**
- Choix du **nombre de groupes**

# Conclusion

- Revue de méthodes pour **regrouper** des patients aux **trajectoires similaires**

Appliqué ici à des **traitements** mais **applicable à n'importe quelles séquences d'états**
(séquence d'hospitalisations, consultations, consommation de soins…)

- **Méthodes exploratoires**, nécessitant une **validation métier** des clusters obtenus

- L'**Optimal Matching** est une méthode prometteuse :

**Simple** à mettre en œuvre (package TraMineR)

Autorise des **distorsions temporelles** (contrairement à la distance de Hamming)

Permet une **flexibilité** dans l'attribution des coûts

Produit des **représentations graphiques** des clusters obtenus

**Pour aller plus loin :**

- **Plusieurs dimensions** dans le calcul des distances (analyse multi-séquences)
- **Prédiction** de l'appartenance aux clusters

# Merci de votre attention

*Questions et remarques sont bienvenues*

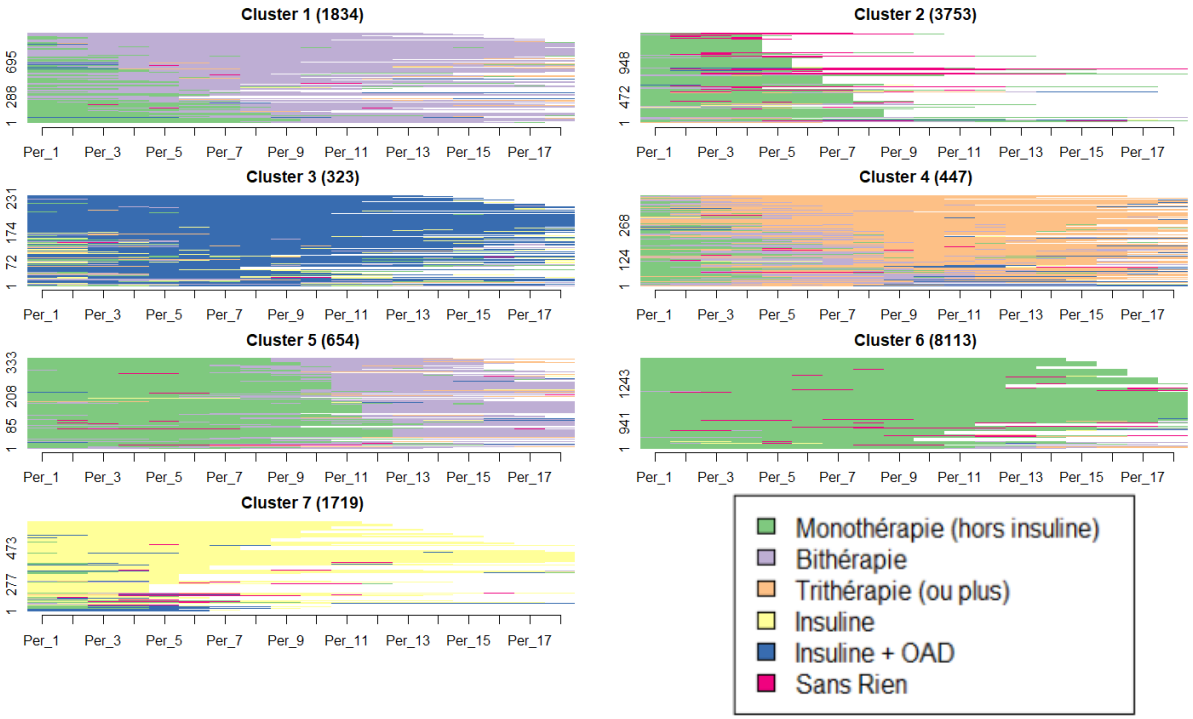
Contacts :
Oriane Bretin oriane.bretin@iqvia.com
Pascale Rondeau pascale.rondeau@iqvia.com

# Résultats et interprétation

*Optimal Matching - 7 groupes - 2 méthodes de calcul de coûts*

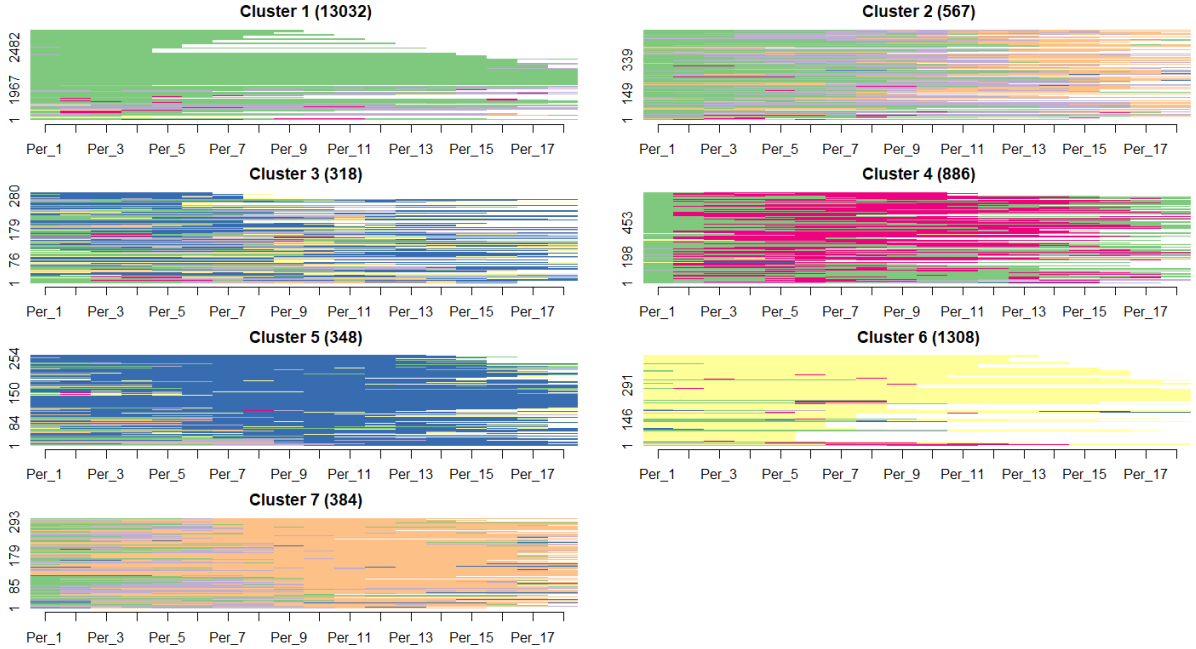
## 1. Coûts basés sur les taux de transition entre états

*Il est peu cher de substituer deux états qui se succèdent souvent dans la population*




## 2. Coûts basés sur les indels

*Les états plus fréquents coûtent plus chers à insérer*




→ **Augmenter le nombre de groupes** peut faire ressortir plus de nuances entre les trajectoires.