

# Utilisation de la PLS-DA pour la caractérisation de profils patients issus d'un clustering TAK sur leur historique de traitement dans le SNDS

## Contexte

Dans le cadre d'une étude observationnelle avec des données en vie réelle des bases du Système National des Données de Santé (SNDS) sur la Dermatite Atopique (DA), une description de l'historique des séquences de traitements a été réalisée. En appliquant un algorithme de clustering non-supervisé prenant en compte la temporalité des traitements, quatre groupes de patients ont été identifiés issus du clustering de la méthode TAK®.

- A** Patients avec de très longues séquences de traitementst avant l'inclusion
- B** Patients avec des séquences de traitements moyennes avant l'inclusion
- C** Patients avec de courtes séquences de traitements avant l'inclusion
- D** Patients avec très peu de traitements avant l'inclusion

## Objectif

L'objectif de cette analyse est de caractériser ces groupes de patients à travers leurs antécédents du parcours de soins en appliquant une méthode multivariée supervisée de réduction de l'espace : l'analyse discriminante des moindres carrés partiels (PLS-DA).

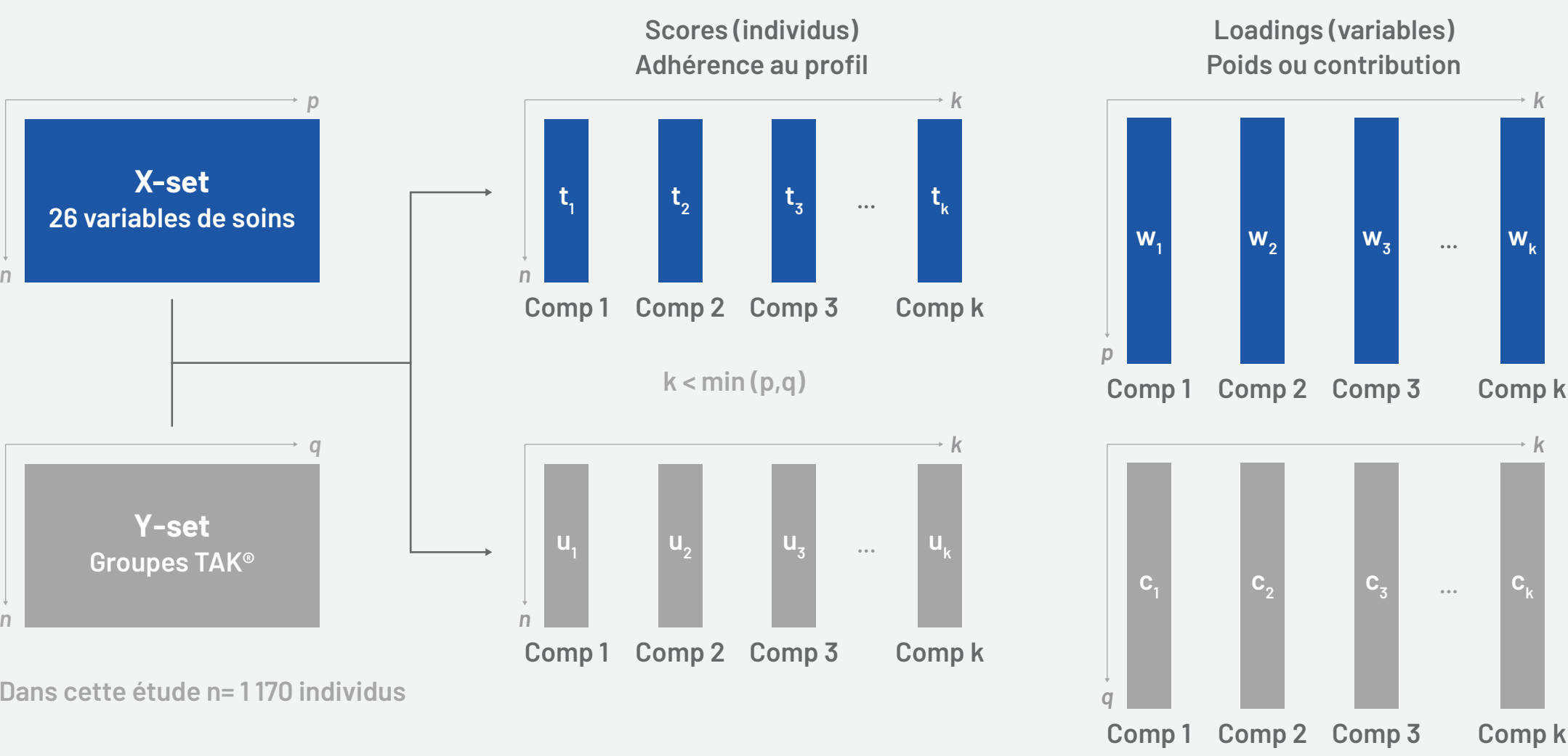
## Méthodes

La PLS-DA est une méthode multivariée d'apprentissage supervisé de réduction de l'espace reposant sur une décomposition PLS et de la classification linéaire.

Elle permet d'identifier des combinaisons des variables de soins et caractéristiques patients c'est à dire des profils permettent de discriminer ou prédire une variable réponse. Pour résumer, un ensemble de variables quantitatives ( $X_1, X_2, \dots, X_k$  : variables explicatives faisant partie du set des prédicteurs, ici un ensemble de 26 variables caractérisant la phase de pré-inclusion comprenant la classe d'âge, le nombre de consultations chez le généraliste, chez le dermatologue, consultations hospitalières, nombre d'hospitalisations, hospitalisations et liés à la dermatite, entre autres) est réduit et résumé avec des composantes PLS construites avec connaissance de la variable réponse catégorielle qualitative Y que l'on cherche à prédire (ici un des quatre groupes **A**, **B**, **C** et **D** issus du clustering TAK® allant de ceux les plus médicamenteux à ceux ayant le moins de traitements avant l'inclusion) codée avec une série d'indicateurs de ses classes (codage disjonctif complet).

La régression PLS-DA produit itérativement une série de facteurs deux à deux orthogonaux (PLS-DA-X et PLS-DA-Y). Ces facteurs sont caractérisés par :

- Des « loadings » : l'ensemble des poids (en lien avec le X-set) est noté  $w_h(p*1)$  en supposant que h composantes soient gardées.
- Des scores : les scores (en lien avec le X-set), notés  $t_h$  sont obtenus en multipliant les variables originales par les poids, tel que  $t_h = Xw_h$ .



## Conclusion

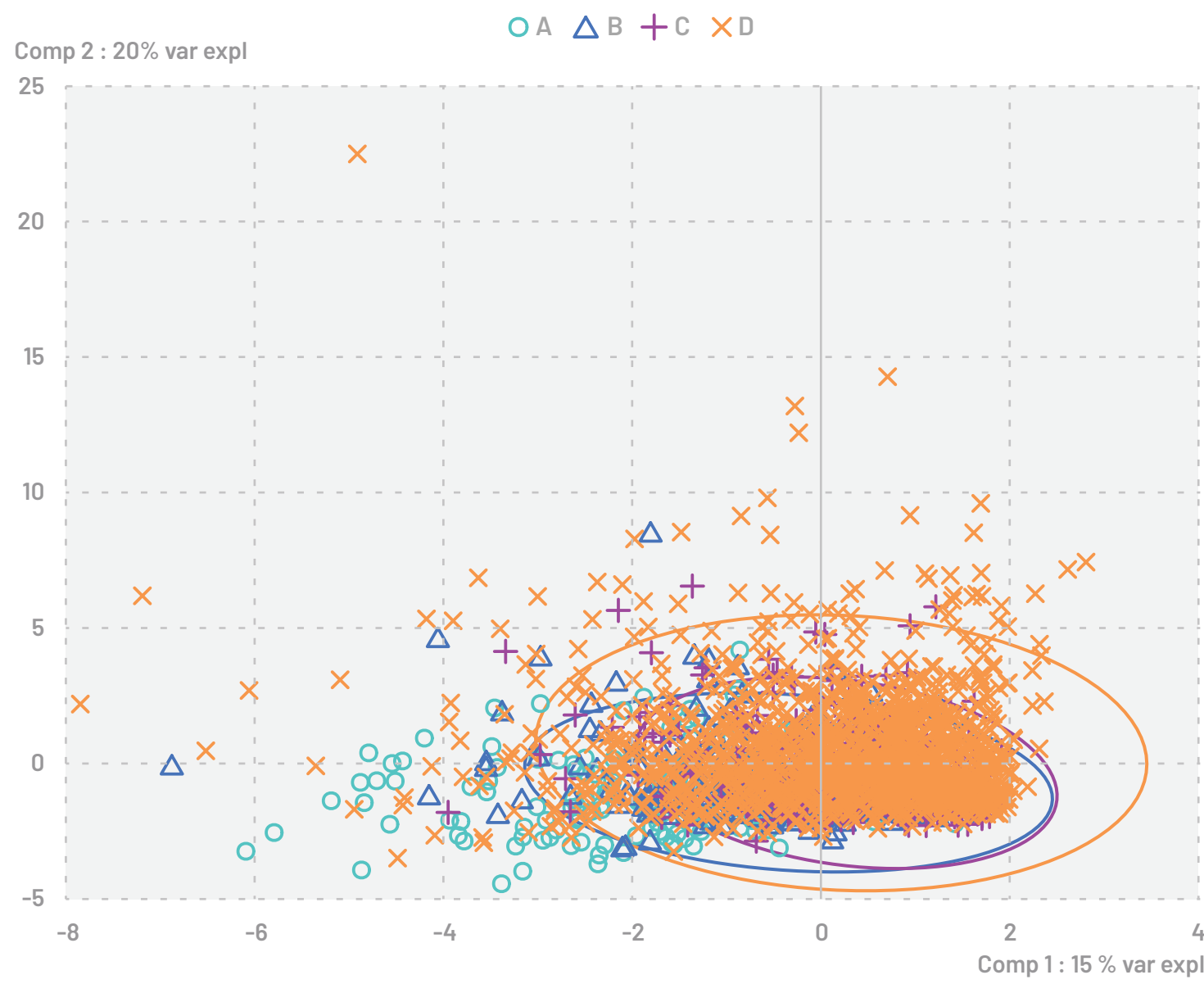
Cette étude a permis d'identifier des signatures de parcours de soins permettant de caractériser des groupes de patients aux trajectoires de traitements différentes. La PLS-DA a été utilisée à des fins de la sélection de caractéristiques et pour ses propriétés discriminatoires, elle pourrait dans un autre contexte être utilisée pour faire des prédictions de classes pour de nouvelles observations.

## Résultats

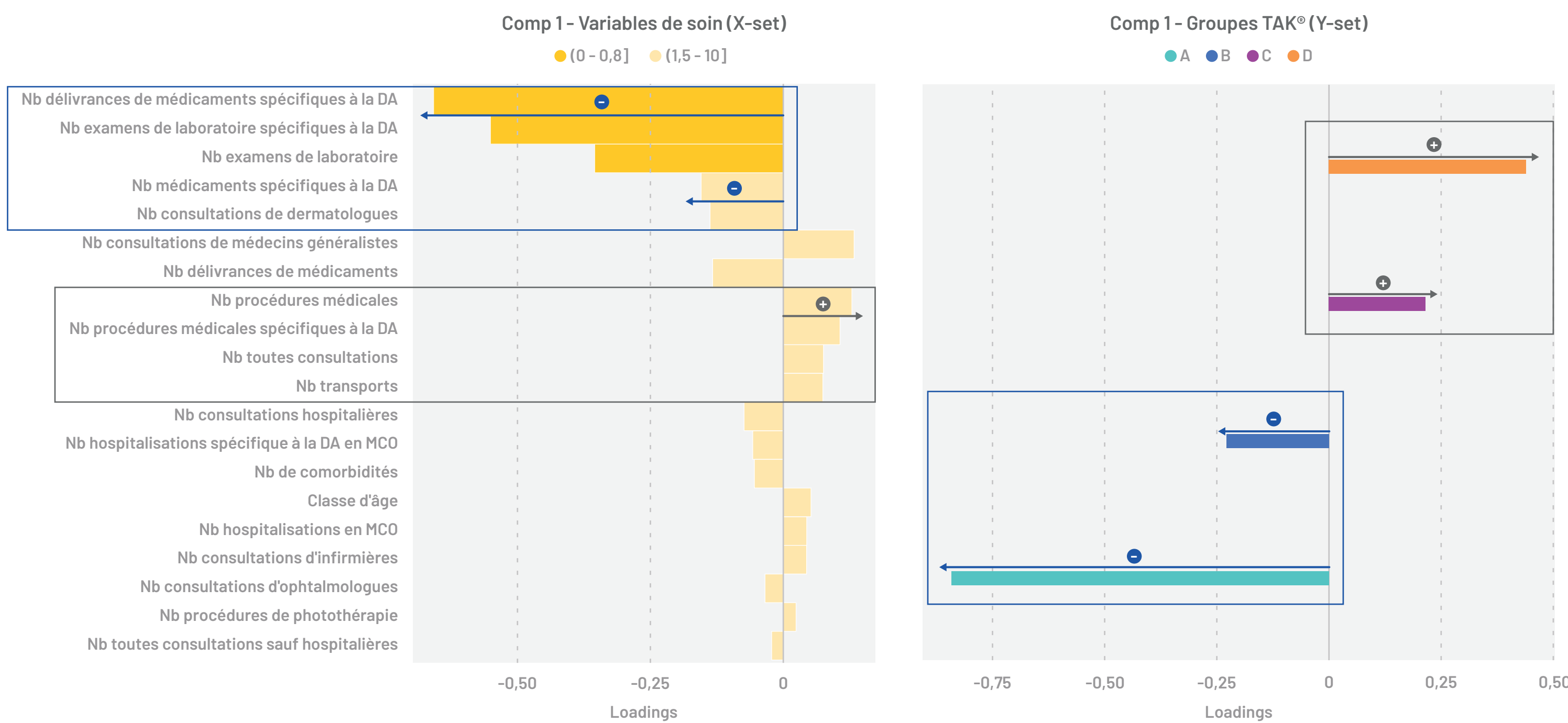
### Profils identifiés

Quatre profils de patients ou composantes PLS-DA ont été identifiés avec une variabilité expliquée allant de 20 % à 3,4 %.

Seul le premier profil (Comp 1) a permis une discrimination entre les groupes de TAK®, en particulier entre patients ayant reçu de longues séquences de traitements précédant l'entrée dans l'étude par rapport aux groupes avec peu de traitements (**A** vs. **C**, **D**).



### Loadings Comp 1



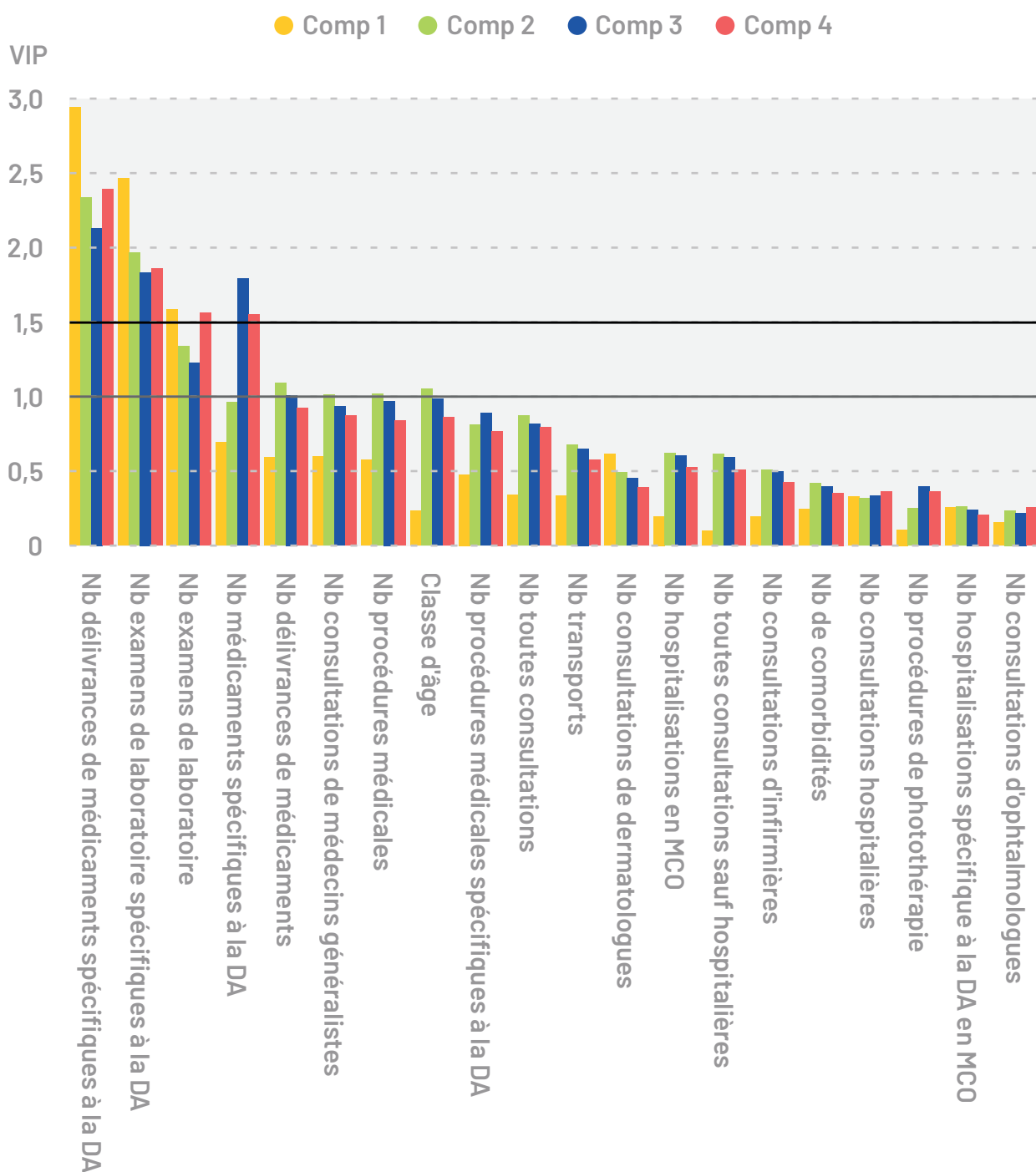
→ Interprétation conjointe basée sur le sens (signe) du loading pour les variables définies et le groupe TAK®

### Variable Importance in the Projection (VIP)

Le VIP, importance de la variable  $x_i$  ( $i=1, \dots, p$ ) pour la prédiction de y dans un modèle à k composantes s'écrit :

$$VIP_{ki} = \sqrt{\frac{p}{\sum_{h=1}^k R^2(y, t_h)} \sum_{h=1}^k R^2(y, t_h) w_{hi}^2}$$
 où  $t_h = Xw_h$  et y la variable réponse

- Estimation quantitative du pouvoir discriminatoire de chaque variable
- Somme pondérée des corrélations au carré entre chaque composante PLS-DA et les variables de soins initiales



### Variables de soins avec VIP > 1,5 sur Comp 1

- Nombres de dispensations de médicaments spécifiques à la DA
- Nombre de tests de médicaments spécifiques à la DA
- Nombre de tests de laboratoire

→ Les valeurs de VIP supérieures au seuil de Wold (0,8) peuvent également être utilisées, bien que dans les données à haute dimensionalité les variables soient jugées pertinentes lorsque  $VIP > 1$  ou 1,5

### Prédiction et courbes ROC

Le taux d'erreur de classification était élevé et ne pouvait permettre la réalisation de prédictions fiables. Cependant, les résultats peuvent être interprétés comme des modèles permettant de mettre en évidence des profils de variables de soins associés aux groupes TAK®, en particulier pour le groupe **A** vs **B**, **C** et **D**.

