Heva

Marie Laurent, Fanny Raguideau, Martin Prodel, Nada Assi, Eléonore Herquelot, Benjamin Grenier

Heva, Lyon, France

39th ICPE

August 23-27, 2023, Halifax, Canada

527

Use of a Clustering Algorithm for Treatment Sequences (TAK®) as a Pre-processing Step to Explanatory and Predictive Models

Background

To summarize the treatment sequences information is complex as it is a longitudinal process involving several treatments and combinations of treatments. The TAK® (Time sequence analysis through K-clustering) is a non-supervised clustering algorithm of time sequences which can be used as a pre-processing step whose results (clusters) summarize treatment sequences information.

Objectives

To describe how to characterize the treatment sequence during the follow-up and to take into account the history of treatment sequences in analyses using the TAK® results.

Methods

TAK®

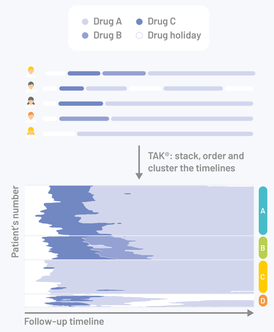
The TAK® is a 3-steps clustering algorithm:

1. Each patient treatment sequence is modeled as a timeline,

2. Timelines are clustered using an Agglomerative Clustering configured with the Hamming distance,

3. Timelines are stacked one on top of the other, cluster by cluster, in the order given by the dendrogram.

Here we focused on the step 2 results to extract a new feature: the cluster in which the patient is classified. This new feature summarizes in one simple categorical variable the treatment sequence of the patient, and can therefore be used in explanatory and predictive models[1].

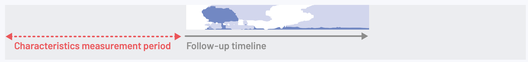


Explanatory models

→ Is the heterogeneity of treatment sequences associated with age, comorbidities, or hospital setting?

When **clusters are used as the outcome in a model** (multinomial logistic regression, multilevel model), one can assess the associations between treatment sequences and patients' characteristics.

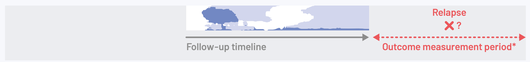
Multinomial logistic regression: Estimates the association between categorical variables and different groups of the TAK clusters.



Predictive model

→ Is the probability of occurrence of an event after the TAK period higher in a given group?

When **clusters are used as a covariable in a model** (eg. Kaplan Meier model, Logistic Regression, Cox model), one can assess the influence of the treatment sequence on overall survival, progression-free survival, or time to adverse event (measured after the period included in the TAK®).



*The measurement period for the outcome can only start after the end of the time period presented in the TAK® and used to determine the groups

Partial Least Squares - Discriminant Analysis (PLS-DA)

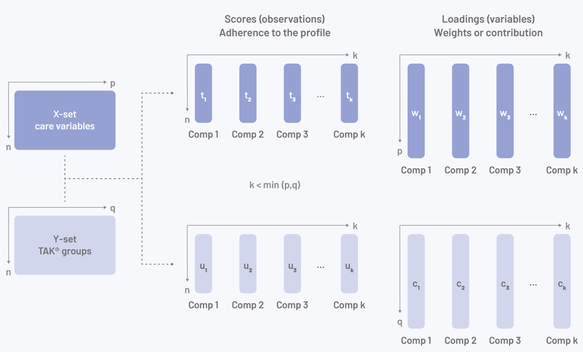
→ How do TAK groups relate to a set of variables?

Group characterization using PLS-DA on course history.

PLS-DA is a supervised multivariate dimension reduction method combining features of PLS decomposition and linear regression[2]. PLS-DA allows the identification of profiles that are combinations of variables, the adherence of patients to these profiles is quantified via their scores on each component. These components are built to provide discrimination or prediction for the outcome of interest. The approach is holistic as it summarizes the relative contributions of all examined variables instead of univariately assessing each one at an equal level of all other variables.

Part 1: Do care variables discriminate between TAK® groups?

Part 2: Which care variables are most important for classification and/or prediction?



Variable Importance in the Projection (VIP) values:

- *Quantitative estimation of the discriminative power of each variable*
- *Weighted sum of squared correlations between PLS-DA component and original explanatory variables*

Results

Patients' description

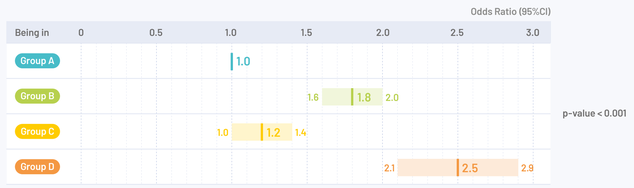
Description of the characteristics and hospital settings of patients from each cluster (age, comorbidities/medical history, year of treatment, region).

→ Patients in cluster D are older than patients in the other clusters

→ There is a higher proportion of patients with cardiovascular disease in clusters C and D

Predictive models

Risk of relapse in the year following the end of the TAK follow-up (logistic regression).



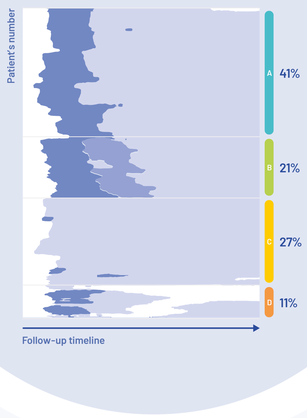
Survival time from the end of the TAK follow-up (Kaplan Meier)

→ The survival rate is better for patients in clusters A and C than for patients in clusters B and D.

→ A Cox proportional hazards model can be used instead of a Kaplan Meier estimator to adjust on patients' characteristics

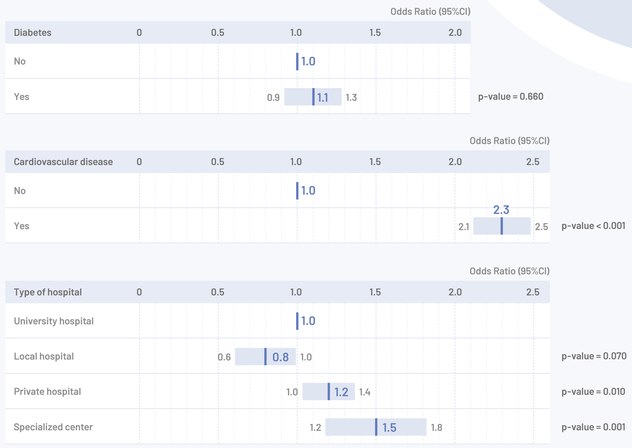
TAK® results

What to do next with the clusters?



Explanatory models

Quantification of the associations between clusters and characteristics/hospital settings of patients, with a multinomial logistic regression model.



→ The different patterns of history of treatment are associated with cardiovascular disease and the type of hospital, but not with diabetes

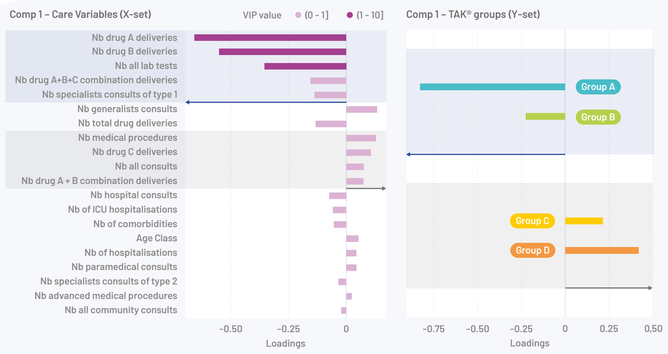
PLS-DA

Identification of care profiles that discriminate between TAK groups.

Loadings Comp 1

→ Joint interpretation based on the direction (sign) of the loading value for the defined variables and each TAK® group

→ VIP values above the Wold threshold (0.8) can also be used, although in high-dimensional data, variables are considered relevant when VIP > 1 or 1.5



→ Subjects belonging to groups A or B have more deliveries of drugs A and B than subjects in groups C and D who in turn have more medical procedures and drug C deliveries.

Conclusion

The clusters obtained from the TAK® are simple to interpret, robust to noise and summarize the information of the patients' treatment sequences (treatments, temporality) into one categorical variable which can be further used in predictive, multivariate and explanatory models.

References

1. Chouaïd C, Grumberg V, Batisse A, Corre R, Giaj Levra M, Gaudin AF, et al. Machine Learning-Based Analysis of Treatment Sequences Typology in Advanced Non-Small-Cell Lung Cancer Long-Term Survivors Treated With Nivolumab. JCO Clinical Cancer Informatics. 2022;(6):e2100108.

2. Assi N, Beisel J, Nosbaum A, Staumont-Sallé D, Foist M, Schmidt A, et al. Utilisation de la PLS-DA pour la caractérisation de profils patients issus d'un clustering TAK® sur leur historique de traitement dans le Système national des données de santé. Revue d'Épidémiologie et de Santé Publique. 2023 May;71:101800.