Check for updates

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
2018, Vol. 27(1) 97–113
© The Author(s) 2015
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280215621591
journals.sagepub.com/home/smm
SAGE

Clustering of longitudinal data by using an extended baseline: A new method for treatment efficacy clustering in longitudinal data

Catherine Schramm,[1,2,3,4] Céline Vial,[5]
Anne-Catherine Bachoud-Lévi[2,4,6] and Sandrine Katsahian[1,7]

Abstract

Heterogeneity in treatment efficacy is a major concern in clinical trials. Clustering may help to identify the treatment responders and the non-responders. In the context of longitudinal cluster analyses, sample size and variability of the times of measurements are the main issues with the current methods. Here, we propose a new two-step method for the Clustering of Longitudinal data by using an Extended Baseline. The first step relies on a piecewise linear mixed model for repeated measurements with a treatment-time interaction. The second step clusters the random predictions and considers several parametric (model-based) and non-parametric (partitioning, ascendant hierarchical clustering) algorithms. A simulation study compares all options of the clustering of longitudinal data by using an extended baseline method with the latent-class mixed model. The clustering of longitudinal data by using an extended baseline method with the two model-based algorithms was the more robust model. The clustering of longitudinal data by using an extended baseline method with all the non-parametric algorithms failed when there were unequal variances of treatment effect between clusters or when the subgroups had unbalanced sample sizes. The latent-class mixed model failed when the between-patients slope variability is high. Two real data sets on neurodegenerative disease and on obesity illustrate the clustering of longitudinal data by using an extended baseline method and show how clustering may help to identify the marker(s) of the treatment response. The application of the clustering of longitudinal data by using an extended baseline method in exploratory analysis as the first stage before setting up stratified designs can provide a better estimation of treatment effect in future clinical trials.

Keywords

Clustering, longitudinal data, personalized medicine, treatment effect, Huntington's disease, obesity

1 Introduction

Heterogeneity in treatment efficacy is one of the biggest concerns in personalized medicine. However, clustering may help to identify the treatment responders and the non-responders. Clustering is an unsupervised learning method that allows a hidden structure to be found in unlabelled data. It relies on an algorithm to minimize within-cluster variability (internal cohesion) and maximize between-cluster variability (external isolation).[1] Here, we

[1]INSERM UMRS1138, Centre de Recherche des Cordeliers, E22, Université Paris Descartes, Université Pierre et Marie Curie, Paris, France
[2]INSERM U955 E01, Neuropsychologie interventionnelle Laboratory IMRB, Créteil, France
[3]Université Pierre et Marie Curie, Paris 6, Paris, France
[4]École Normale Supérieure, Institut d'Études de la Cognition, Paris, France
[5]Université de Lyon, CNRS UMR 5208, Polytech Lyon-Université de Lyon 1, Institut Camille Jordan, Villeurbanne, France
[6]Assistance Publique-Hôpitaux de Paris, National Reference Center for Huntington's Disease Henri Mondor Hospital, Créteil, France
[7]Assistance Publique-Hôpitaux de Paris, Service d'informatique et statistiques, Georges Pompidou European Hospital, Paris, France

Corresponding author:
Catherine Schramm, Centre de Recherche des Cordeliers, Escalier D, 1er étage 15 rue de l'école de médecine 75006 Paris, France.
Email: cath.schramm@gmail.com

explore the specific context of a rare and progressive disease with a small sample size and a treatment effect that is measured longitudinally. Thus, the information provided by the longitudinal data is not restricted to a single value but must consider the entire trajectory of a continuous score. More precisely, we focus on the change of slope after the treatment initiation. For this longitudinal cluster analysis, the data may be obtained from cohort studies or from clinical trials (treatment arm) with the treatment initiation during the follow-up. In both cases, repeated measurements of the patients' scores are recorded before and after the initiation of treatment.

The current parametric and non-parametric methods for longitudinal cluster analysis are being increasingly used in medical research.[2–6] Parametric methods relate to mixture modeling techniques, in particular through latent-class mixed models (LCMM). They assess the influence of latent growth trajectory class membership on the outcome to highlight the distinct patterns of evolution.[7] The mixed model allows the within-subject correlation and the variability of the outcome trajectory between subjects to be taken into account. Latent-class is defined by using the assumption of a mixture of Gaussian distributions for the random effects.[8] Clusters and model parameters are estimated simultaneously. The main advantage of these parametric methods is that the usual statistical tests and inferences can be performed; however, if they are to be efficient these methods often require a large sample of patients, which might not be the case for rare diseases or innovative therapies like cell transplant or gene therapy. Non-parametric methods relate to classical algorithmic approaches such as K-Means for Longitudinal data (KML).[9] Such algorithms consider the distance between patients' score rather than the shape of the evolution which does not address the initial problem of the change in the slope due to treatment effect. Furthermore, they need a constant measurement delay between patients, which is not a reasonable assumption in cohort studies. The limits of these methods suggest the need for a new method to cluster longitudinal data according to treatment effect when there is both a small sample and variability in the times of measurement. We propose a Clustering of Longitudinal data with an Extended Baseline (CLEB) method. This new method comprises two steps: first, building a linear mixed model with an extended baseline and second clustering the random predictions through a model-based algorithm. However, other strategies of clustering could be planned in the second step, notably non-parametric algorithms.

The objectives of this paper are (i) to present this new method and (ii) to compare the model-based and non-parametric strategies. The CLEB method is described in section CLEB. The simulation study settings and results are presented in, respectively, sections **The simulation study procedure** and **Results of the simulation study**. The different strategies of the CLEB method are evaluated and compared to the LCMM algorithm. The CLEB method is illustrated in section Applications with a real data set of Huntington's disease patients and a real data set of women suffering from obesity. The results are discussed in section Discussion.

2 CLEB

The CLEB method clusters patients according to treatment effect. The two steps of this method are described in this section.

Consider data from $i = 1, ., N$ patients in a longitudinal study assessed n_i times with y_{ij} the jth outcome measure of patient i at time t_{ij}. Patients initiate their treatment during the follow-up and the times are realigned such that 0 corresponds to the time of treatment initiation. Thus, there are negative times ($t_{ij} < 0$) for measurements before the treatment initiation. The lag $\tau \geq 0$ between treatment initiation and treatment effect is used to define two phases: the baseline phase for $t_{ij} < \tau$ and the treatment effect phase for $t_{ij} \geq \tau$.

2.1 Step 1: The mixed model with extended baseline for treatment-time interaction

Figure 1 represents the CLEB algorithm.

The basic assumption of the method is that the individual's responses vary linearly according to phase. The polynomial model with extended baseline[10] could be adapted to the data as follows

$$y_{ij} = a_{0i} + a_{1i} \times t_{ij} + a_{2i} \times (t_{ij} - \tau) \times 1(t_{ij} \geq \tau) + \varepsilon_{ij} \tag{1}$$

where $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

The random effect model for longitudinal data[11] takes into account all the available information, allowing the model to deal with missing data and numbers and times of measurements that are not identical.[12] It takes into account both within- and between-patient variability. Assuming between-patient variability in the baseline phase

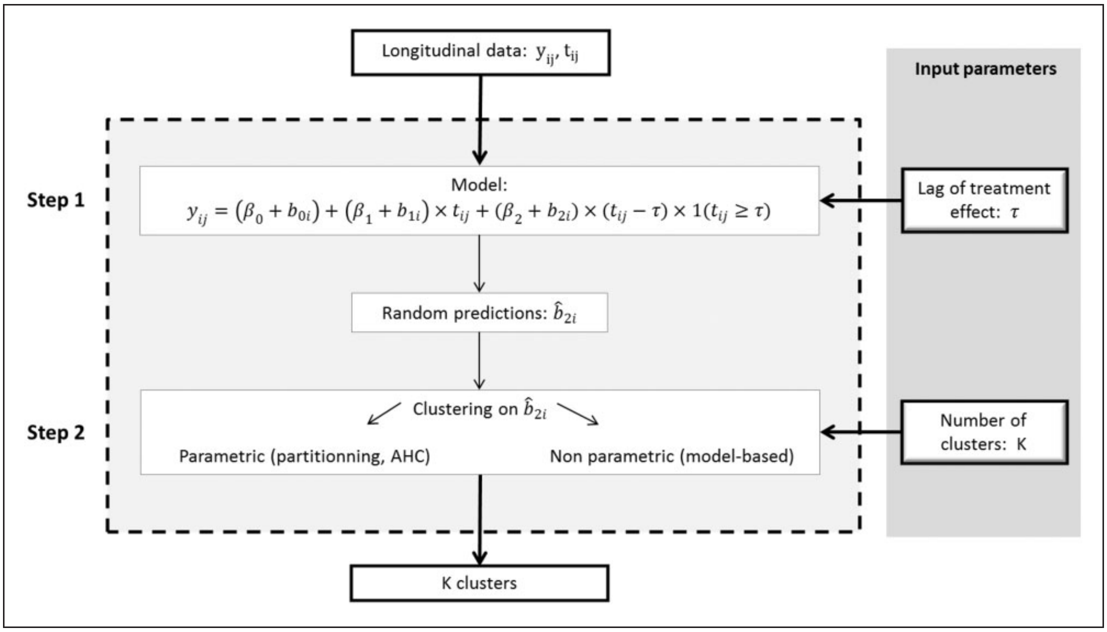

Longitudinal data: y_{ij}, t_{ij}
Input parameters
Step 1
Model:
$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) \times t_{ij} + (\beta_2 + b_{2i}) \times (t_{ij} - \tau) \times 1(t_{ij} \geq \tau)$
Lag of treatment effect: τ
Random predictions: $\hat{b}_{2i}$
Step 2
Clustering on $\hat{b}_{2i}$
Parametric (partitionning, AHC)
Non parametric (model-based)
Number of clusters: K
K clusters

Figure 1. General architecture of the CLEB method.

(intercept and slope) and in the treatment effect phase, $\boldsymbol{a_i}$ could be decomposed as the sum of a fixed effect $\boldsymbol{\beta}$ and a random effect $\boldsymbol{b_i} \sim \mathcal{N}(0, \boldsymbol{D})$ with $\boldsymbol{D} = \begin{pmatrix} \sigma_0^2 & 0 & 0 \\ 0 & \sigma_1^2 & 0 \\ 0 & 0 & \sigma_2^2 \end{pmatrix}$ such that equation (1) becomes

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) \times t_{ij} + (\beta_2 + b_{2i}) \times (t_{ij} - \tau) \times 1(t_{ij} \geq \tau) + \varepsilon_{ij} \tag{2}$$

The mean score at treatment initiation is β_0. The mean progression slope of y_{ij} during the baseline phase is β_1 and the mean progression slope of y_{ij} during the treatment effect phase is the sum $\beta_1 + \beta_2$, where β_2 is the fixed estimate associated with the treatment effect. The third component of equation (2) can be considered as a time-treatment interaction.

The estimation of parameters is made using restricted maximum likelihood, but only the predictions of the random parameters ($\hat{\boldsymbol{b}}_0$, $\hat{\boldsymbol{b}}_1$ and $\hat{\boldsymbol{b}}_2$) are collected. They will be used in the second step of the CLEB method.

2.2 Step 2: Clustering on random predictions

The distribution of $\hat{\boldsymbol{b}}_2$ is assumed to be a mixture of K Gaussian distributions, each distribution corresponding to one cluster.[8] For two clusters A and B of treatment effect $\hat{b}_{2i} \sim p\mathcal{N}(\mu_{2,A}, \sigma^2_{2,A}) + (1-p)\mathcal{N}(\mu_{2,B}, \sigma^2_{2,B})$, where p (resp. $1-p$) is the proportion of subjects in cluster A (resp. B), $\mu_{2,A}$ (resp. $\mu_{2,B}$) and $\sigma^2_{2,A}$ (resp. $\sigma^2_{2,B}$) are the mean and variance of the individual treatment effect ($\hat{b}_{2i}$) of the patients in cluster A (resp. B). The second step of the CLEB method consists of estimating the mixture of distributions using an Expectation-Maximization (EM) model-based algorithm. EM model-based algorithm and non-parametric algorithms are presented in the following subsections.

2.2.1 EM model-based algorithm

Let us now briefly describe the model-based algorithm to cluster patients using $\hat{\boldsymbol{b}}_2$. This parametric model supposes a Gaussian distribution of $\hat{\boldsymbol{b}}_2$ for each cluster.[13] Let f be the density function of the mixture defined by

$$f(\hat{\boldsymbol{b}}_2) = \sum_{k=1}^{K} \boldsymbol{\pi_k} \Phi(\hat{\boldsymbol{b}}_2 | \mu_{2,k}, \sigma^2_{2,k}) \tag{3}$$

where $\boldsymbol{\pi_k}$ is the probability that a subject belongs to the cluster k and $\Phi(\hat{\boldsymbol{b}}_{\boldsymbol{2}}|\mu_{2,k},\sigma^2_{2,k})$ is the density function from the distribution $\mathcal{N}(\mu_{2,k},\sigma^2_{2,k})$. The method uses the maximization of likelihood in the EM algorithm to estimate $\mu_{2,k}$ and $\sigma^2_{2,k}$ for each cluster k and $\pi_{k,i}$ for each cluster k and each patient i.[14] Two parameterizations could be considered:

- E parameterization: equal variance between clusters
- V parameterization: variable variances between clusters

Then $\pi_{k,i}$ is used to classify each patient in the cluster with the higher probability such that $i \in k'$ if $\pi_{k',i} = \max_k \pi_{k,i}$.

This algorithm could be extended to the case of clustering on the multivariate component $\boldsymbol{\theta} = (\hat{\boldsymbol{b}}_{\boldsymbol{0}},\hat{\boldsymbol{b}}_{\boldsymbol{1}})$ or $\boldsymbol{\theta} = (\hat{\boldsymbol{b}}_{\boldsymbol{1}},\hat{\boldsymbol{b}}_{\boldsymbol{2}})$ or $\boldsymbol{\theta} = (\hat{\boldsymbol{b}}_{\boldsymbol{0}},\hat{\boldsymbol{b}}_{\boldsymbol{1}},\hat{\boldsymbol{b}}_{\boldsymbol{2}})$. In this case, the density function g of the multivariate mixture is defined by

$$g(\boldsymbol{\theta}) = \sum_{k=1}^{K} \boldsymbol{\pi_k}\Phi(\boldsymbol{\theta}|\boldsymbol{\mu_k},\boldsymbol{\Sigma_k}) \tag{4}$$

where $\Phi(\boldsymbol{\theta}|\boldsymbol{\mu_k},\boldsymbol{\Sigma_k})$ is the density function from the multivariate normal distribution $\mathcal{N}_{dim(\boldsymbol{\theta})}(\boldsymbol{\mu_k},\boldsymbol{\Sigma_k})$ with $\boldsymbol{\mu_k}$ the vector of means and $\boldsymbol{\Sigma_k}$ the variance−covariance matrix.

In the multivariate case, ten parameterizations may be considered. They concern the variance−covariance matrix structure defined according to three geometric parameters: volume (equal: E or variable: V), shape (equal: E, variable: V, identity: I) and orientation (equal: E, variable: V, identity: I) between clusters.[15]

2.2.2 *Non-parametric alternatives (k-means, k-medoids, agglomerative hierarchical clustering)*

The k-means[16,17] and the k-medoids[18] are iterative algorithms partitioning the data space into Voronoi cells. They attribute data points to clusters by minimizing the distance between each point and the mean of the cluster (in k-means) or the most central value (in k-medoids).

Agglomerative Hierarchical Clustering (AHC) procedures[19] agglomerate the data from N single-member clusters into one cluster containing all data points and stop when the expected number of clusters is reached. The AHC-single linkage, -complete linkage and -average linkage define the distance between clusters as, respectively, the minimal, maximal and average distance between the data points of each cluster.

Several definitions for calculating the distances between two data points exist. Those considered in the simulation study are defined below. Let $\boldsymbol{\theta}$ be the vector of random predictions such that $\boldsymbol{\theta} = \hat{\boldsymbol{b}}_{\boldsymbol{2}}$ or $\boldsymbol{\theta} = (\hat{\boldsymbol{b}}_{\boldsymbol{0}},\hat{\boldsymbol{b}}_{\boldsymbol{1}})$ or $\boldsymbol{\theta} = (\hat{\boldsymbol{b}}_{\boldsymbol{1}},\hat{\boldsymbol{b}}_{\boldsymbol{2}})$ or $\boldsymbol{\theta} = (\hat{\boldsymbol{b}}_{\boldsymbol{0}},\hat{\boldsymbol{b}}_{\boldsymbol{1}},\hat{\boldsymbol{b}}_{\boldsymbol{2}})$. The Euclidean, Canberra, Manhattan, Maximum, Pearson, and Correlation distances for two patients i and j are defined as

$$\begin{aligned}
d_{\text{Euclidean}}(i,j) &= \sqrt{\sum_{\hat{\boldsymbol{b}}_\ell \in \boldsymbol{\theta}} (\hat{b}_{\ell,i} - \hat{b}_{\ell,j})^2} \\
d_{\text{Canberra}}(i,j) &= \sum_{\hat{\boldsymbol{b}}_\ell \in \boldsymbol{\theta}} \frac{|\hat{b}_{\ell,i} - \hat{b}_{\ell,j}|}{|\hat{b}_{\ell,i}| + |\hat{b}_{\ell,j}|} \\
d_{\text{Manhattan}}(i,j) &= \sum_{\hat{\boldsymbol{b}}_\ell \in \boldsymbol{\theta}} |\hat{b}_{\ell,i} - \hat{b}_{\ell,j}| \\
d_{\text{Maximum}}(i,j) &= \max_{\hat{\boldsymbol{b}}_\ell \in \boldsymbol{\theta}} |\hat{b}_{\ell,i} - \hat{b}_{\ell,j}| \\
d_{\text{Pearson}}(i,j) &= \frac{\sum_{\hat{\boldsymbol{b}}_\ell \in \boldsymbol{\theta}} \hat{b}_{\ell,i}\hat{b}_{\ell,j}}{\sqrt{\sum_{\hat{\boldsymbol{b}}_\ell \in \boldsymbol{\theta}} \hat{b}^2_{\ell,i} \sum_{\hat{\boldsymbol{b}}_\ell \in \boldsymbol{\theta}} \hat{b}^2_{\ell,j}}} \\
d_{\text{Correlation}}(i,j) &= \frac{Cov(\boldsymbol{\theta_i},\boldsymbol{\theta_j})}{\sqrt{Var(\boldsymbol{\theta_i})Var(\boldsymbol{\theta_j})}}
\end{aligned}$$

Note that for $\boldsymbol{\theta} = \hat{\boldsymbol{b}}_{\boldsymbol{2}}$, the Pearson distance is always equal to one, the correlation distance is not defined and the Manhattan and Maximum distances are equal to the Euclidean distance. Thus only the Euclidean and Canberra distances will be used in the univariate case.

3 The simulation study procedure

Several scenarios with two subgroups of patients (N_A subjects with a beneficial treatment effect in group A and N_B subjects with a detrimental treatment effect in group B) were considered. The notation is similar to that in section CLEB. Each patient i has at least one visit before the treatment initiation, one visit at treatment initiation and one visit after treatment initiation. The period between visits ($\boldsymbol{d}$) is around one year: $d_{ij} \sim \mathcal{N}(1, \sigma_d^2)$ such that $t_{ij} - t_{ij-1} = d_{ij}$. The outcome for each visit is generated by

$$y_{ij} = \begin{cases} \beta_0^{(A)} + \beta_1^{(A)} \times t_{ij} + \beta_2^{(A)} \times (t_{ij} - \tau_i) \times 1(t_{ij} \geq \tau_i) + \varepsilon_{ij} & \text{if} \quad i \in A \\ \beta_0^{(B)} + \beta_1^{(B)} \times t_{ij} + \beta_2^{(B)} \times (t_{ij} - \tau_i) \times 1(t_{ij} \geq \tau_i) + \varepsilon_{ij} & \text{if} \quad i \in B \end{cases} \tag{5}$$

with $\tau_i \sim \mathcal{N}(1/12, \sigma_\tau^2)$ meaning that the treatment effect is supposed to appear at one month. For $\ell \in \{0, 1, 2\}$ and $x \in \{A, B\}$, $\beta_{\ell i}^{(x)} \sim \mathcal{N}(\mu_\ell^{(x)}, \sigma_\ell^{2(x)})$. The strength of treatment effect is $|\mu_2^{(A)} - \mu_2^{(B)}|$, and the within-treatment variation in group A (respectively B) is $\sigma_2^{(A)}$ (respectively $\sigma_2^{(B)}$). The within-patient variability is generated by $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. The simple case scenario is parameterized with 25 patients per group ($N_A = N_B = 25$), each having 3 to 5 visits before treatment initiation and 3 to 5 visits after treatment initiation, null variance on the period between visits and the lag in treatment effect ($\sigma_d = \sigma_\tau = 0$) and low within-patient variability ($\sigma_\varepsilon = 3$). The mean of the effects in group A ($\mu_0^{(A)}, \mu_1^{(A)}, \mu_2^{(A)}$) is the vector (45, 5, − 5), whereas the mean of the effects in group B ($\mu_0^{(B)}, \mu_1^{(B)}, \mu_2^{(B)}$) is the vector (45, 5, 5). Finally, the between-patient variabilities in groups A ($\sigma_0^{(A)}, \sigma_1^{(A)}, \sigma_2^{(A)}$) and B ($\sigma_0^{(B)}, \sigma_1^{(B)}, \sigma_2^{(B)}$) are both initialized by the vector (5, 1, 1).

The baseline mean parameters were initialized according to the pre-randomization period in a longitudinal clinical trial assessing graft in Huntington's disease (NCT00190450), for which the CLEB method was developed. However, this clinical trial has not yet been published, and the data could not be used as an illustrated example in section Applications.

The simulation study compares the CLEB method, with the different strategies, to the LCMM method. The LCMM method assumes that each cluster, also called the latent-class, is characterized by a specific trajectory modelled by a specific linear mixed model. Both the latent-class membership and the trajectory are explained using covariates. Here, the parameterization of the LCMM method for the trajectory was performed in the same way as in equation (2). Only the third term, that corresponding to the time-treatment interaction was used in the parameterization of the latent-class membership. Likelihood maximization and the EM algorithm were used to estimate the class membership probability and the model parameters simultaneously.

All results are expressed as the mean of the percentage of correctly classified patients among the 1000 databases generated for each scenario.

The simulations and computations were performed using the R software.[20] The CLEB method was performed using the `nlme`[21] package for a mixed model for step 1, the `mclust`[22] package for model-based clustering for step 2, and the `amap`[23] and `cluster`[24] packages for non-parametric algorithms for step 2. The LCMM method was performed using the `lcmm`[25] package.

4 Results of the simulation study

Only the most efficient strategies are presented. Thus, for the CLEB method, the AHC strategies are not reported, and k-means were preferred to k-medoids. For the same reason, except for the specific scenarios in which the treatment effect was influenced by the baseline parameters, only univariate strategies are presented.

Figure 2 displays the percentage of correctly classified patients according to the strength of treatment effect and sample size. When there was a great difference between the simulated treatment effects for the two groups, all strategies allocated almost 100% of patients to the correct cluster. In contrast, when there was no difference between the two groups, all strategies randomly allocated patients to each cluster (50% of correctly classified patients). The CLEB method gave better results than the LCMM regardless of the strategy that was used. The sample size had a greater impact for the CLEB method with the model-based strategies, with a better classification for a larger sample size.

Figure 3 displays the percentage of correctly classified patients according to the natural disease progression variability (variability of slope during the baseline phase: σ_1) and the within-treatment variability (variability of slope change: σ_2). Whatever the strategy, the CLEB method was not affected by the natural disease progression variability, whereas the

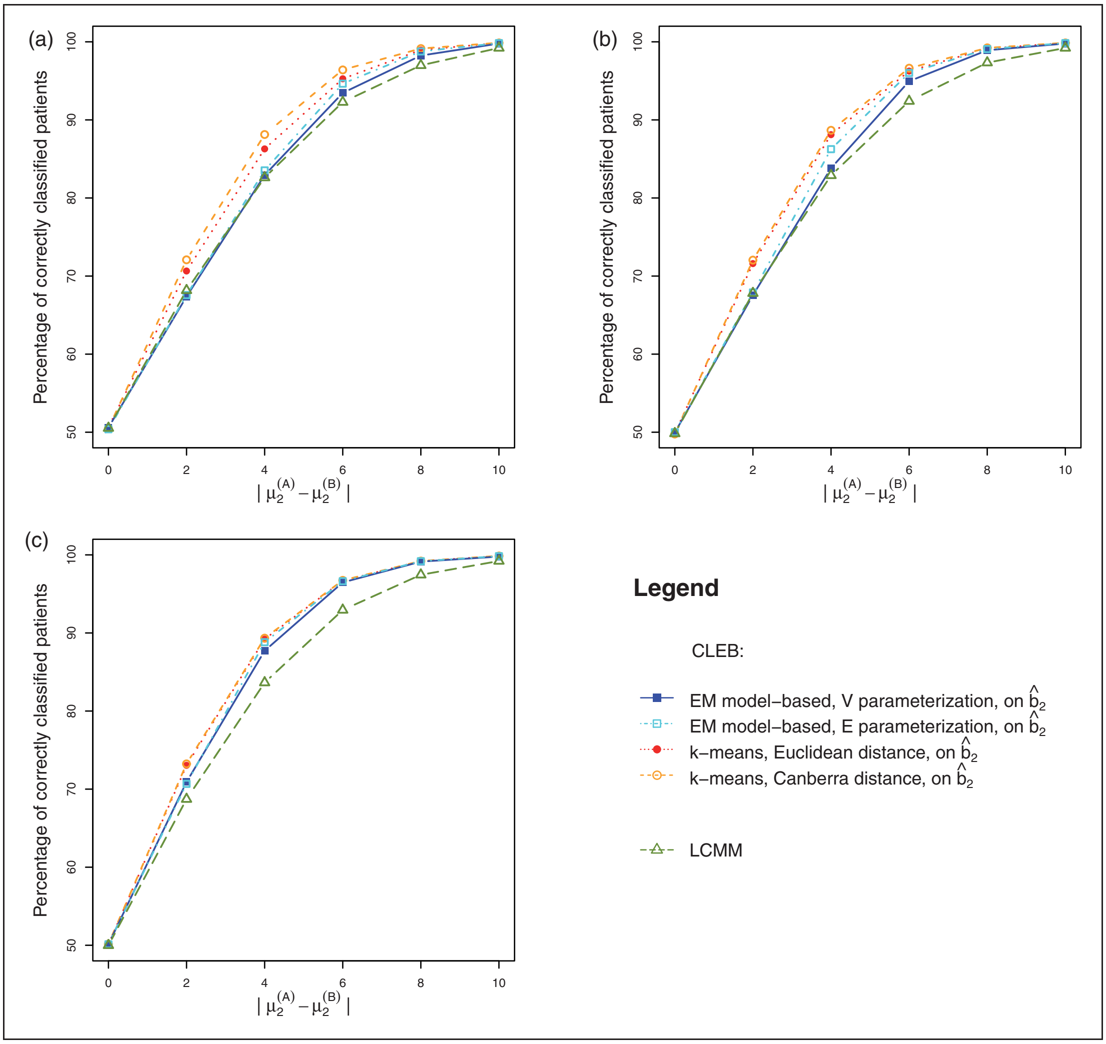


Figure 2. Impact of the strength of treatment effect and the sample size on the percentage of correctly classified patients. (a) $N_A = N_B = 10$; (b) $N_A = N_B = 25$ and (c) $N_A = N_B = 100$.

performances of the LCMM method were worse when σ_1 was greater (Figure 3(a)). Indeed, in LCMM, the subgroup identification and the estimation of the parameters were made simultaneously, leading to a greater influence of the baseline slope on the subgroup definition. The more σ_2 increased, the worse the performances of all methods (Figure 3(b) and (c)). As expected, in CLEB with the EM model-based algorithm, the V parameterization (variable variances between clusters) suffered less of an impact from high variance than the E parameterization (equal variances between clusters) by high variance when the variances were unequal, whereas the V parameterization was equivalent to the E parameterization in the case of equal variances.

Figure 4(a) displays the percentage of correctly classified patients according to the number of subjects in each subgroup. All methods had good performance when the groups were balanced. However, the CLEB method with k-means strategy and Canberra distance did not perform well in the case of unbalanced groups. The LCMM method and the CLEB method with k-means and Euclidean distance were only unable to perform well in an extremely unbalanced case ($N_A = 2$). The CLEB method with EM model-based algorithms was the only model that was not affected at all by unbalanced groups. In the case of $N_A = 2$, the CLEB method with EM model-based algorithms found a small subgroup of responders (Figure 4(b)). All methods have good sensitivity, but the CLEB method with the two EM model-based algorithms has the better specificity.

Figure 5 displays the percentage of correctly classified patients according to the number of time points. For low variability of natural disease progression, an increase in the number of time points after treatment initiation

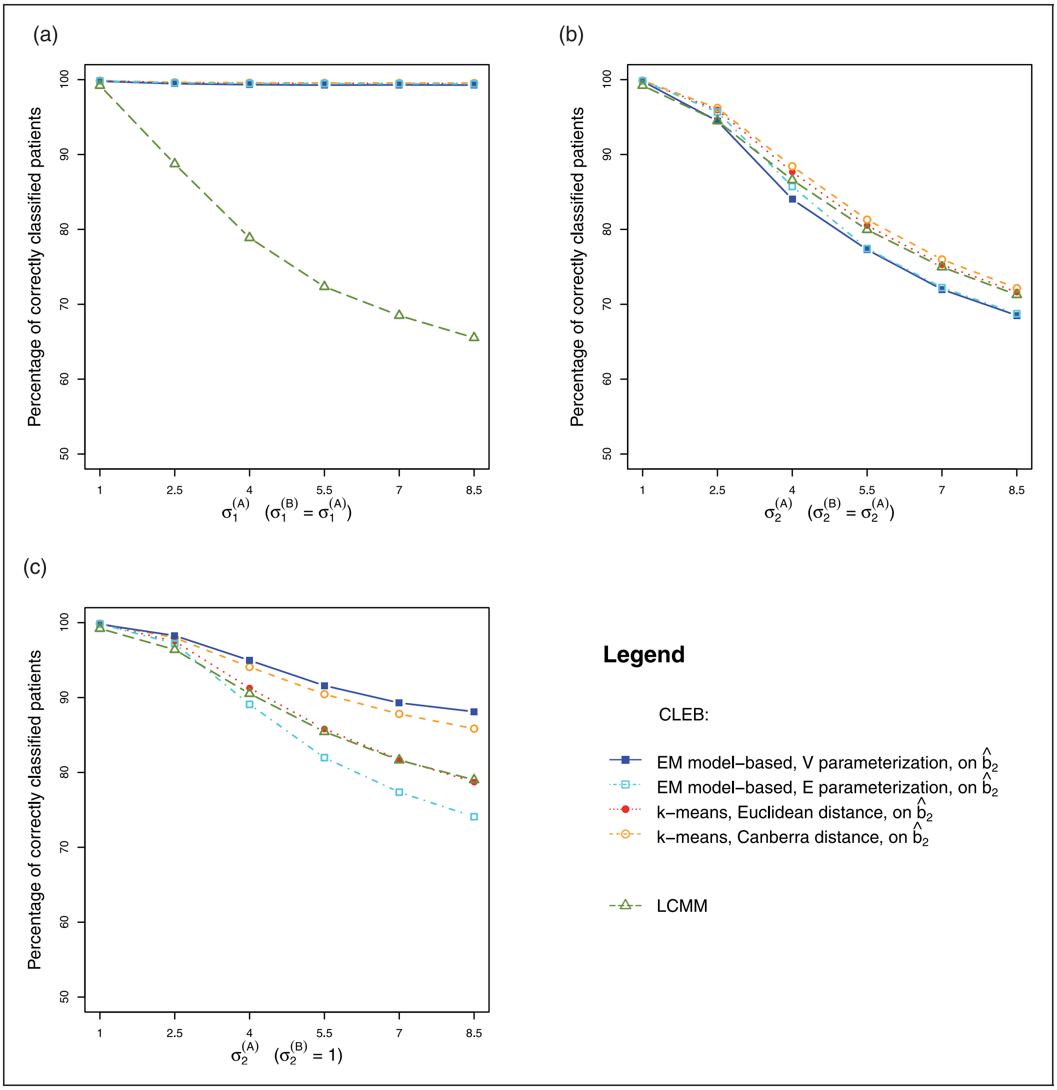
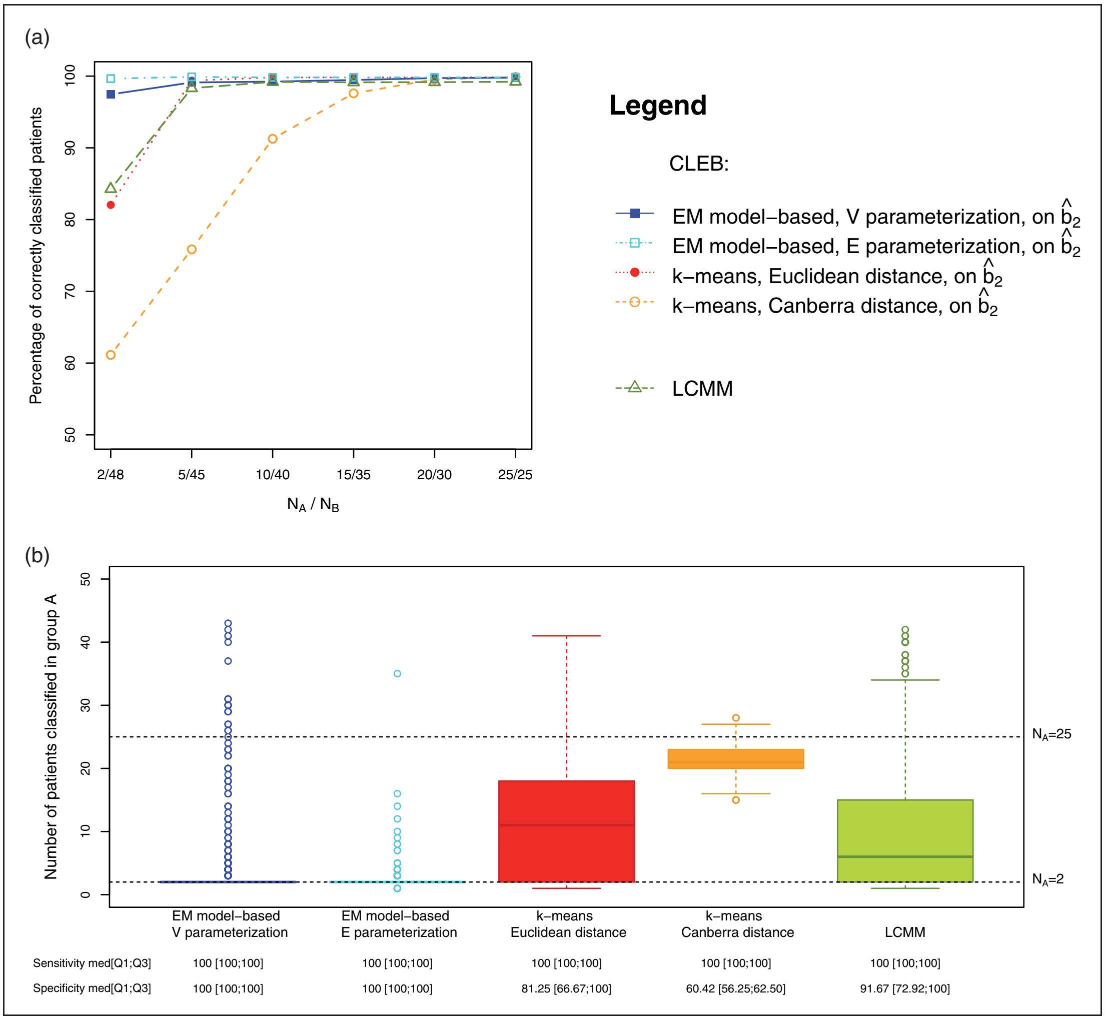


Figure 3. Impact of heterogeneity of the natural disease progression and the within-treatment variation on the percentage of correctly classified patients. (a) Heterogeneity of natural disease progression; (b) Within-treatment variation in two groups and (c) Within-treatment variartion in responders.

improved the performance of all methods, whereas an increase in the number of time points before treatment initiation did not have an impact on their performances. For high variability of natural disease progression, an increase in the number of time points after and/or before treatment initiation improved the performance of the CLEB method, regardless of the associated strategy. The LCMM method, for which variability of natural disease progression had a real impact (Figure 3), did not have an improved performance, regardless of the number of time points. These results suggest that clusters of treatment effects could be identified using only the treatment effect phase in the case of a homogeneous natural disease progression. However, in the case of a heterogeneous natural disease progression, the slope of the baseline phase is necessary to evaluate the treatment effect on the slope change between the baseline and the treatment effect phases.

Figure 6 displays the percentage of correctly classified patients when the baseline characteristics differ in the two clusters. For parametric algorithms, only the better univariate and multivariate strategies are presented; these correspond to, respectively, the V and VVI parameterizations. It should be noted that V corresponds to a hypothesis of variable variance and VVI to a hypothesis of variable variance according to clusters and according to random terms without correlation between clusters. For the non-parametric algorithms, only the k-means with Euclidean distance is presented. Other non-parametric algorithms show similar results.



	EM model–based V parameterization	EM model–based E parameterization	k–means Euclidean distance	k–means Canberra distance	LCMM
Sensitivity med[Q1;Q3]	100 [100;100]	100 [100;100]	100 [100;100]	100 [100;100]	100 [100;100]
Specificity med[Q1;Q3]	100 [100;100]	100 [100;100]	81.25 [66.67;100]	60.42 [56.25;62.50]	91.67 [72.92;100]

Figure 4. The case of unbalanced clusters. (a) Impact on classification and (b) Specific case of N_A = 2 and N_B = 48.

The multivariate strategies improve when the correlation between baseline and treatment effect increases, because $\hat{b}_0$ and $\hat{b}_1$ also capture some useful information for clustering. However, if variability of the baseline parameters is high, the use of these parameters will lead to a less robust classification than the univariate strategies (Figure 6(c) and (f)). The univariate strategy may also be improved by an increase in the difference between the natural slope of the disease in the two subgroups (Figure 6(b) and (e)). Indeed, the difference in natural evolution increases the difference in the slope after treatment initiation, leading to a better classification.

Figure 7 displays the percentage of correctly classified patients in the case of missing data. When the methods were applied using the whole data set, patients with missing data were allocated randomly into clusters (Figure 7(c)), whereas almost 100% of the patients without missing data were allocated to the correct cluster (Figure 7(d)). Only the CLEB method with the EM model-based algorithm and V parameterization was slightly affected by a high rate of missing data. However, applying the method only to subjects without missing data (the complete case study) led to the best results (Figure 7(b)).

The simulation study showed that the CLEB method performs better than LCMM when there is high slope variability. In the CLEB method, univariate strategies were preferred to multivariate strategies. The k-means with Canberra distance was hugely affected by unbalanced groups, to the extent that it was not a reliable strategy. The EM model-based algorithm with V parameterization (the hypothesis of variable variance between clusters) must be the preferred strategy, but the CLEB method with k-means algorithm and Euclidean distance could be more robust when there is a small sample size and low variance. Figure 8 sums up all the strategies considered and shows how the simulation study reached this conclusion.

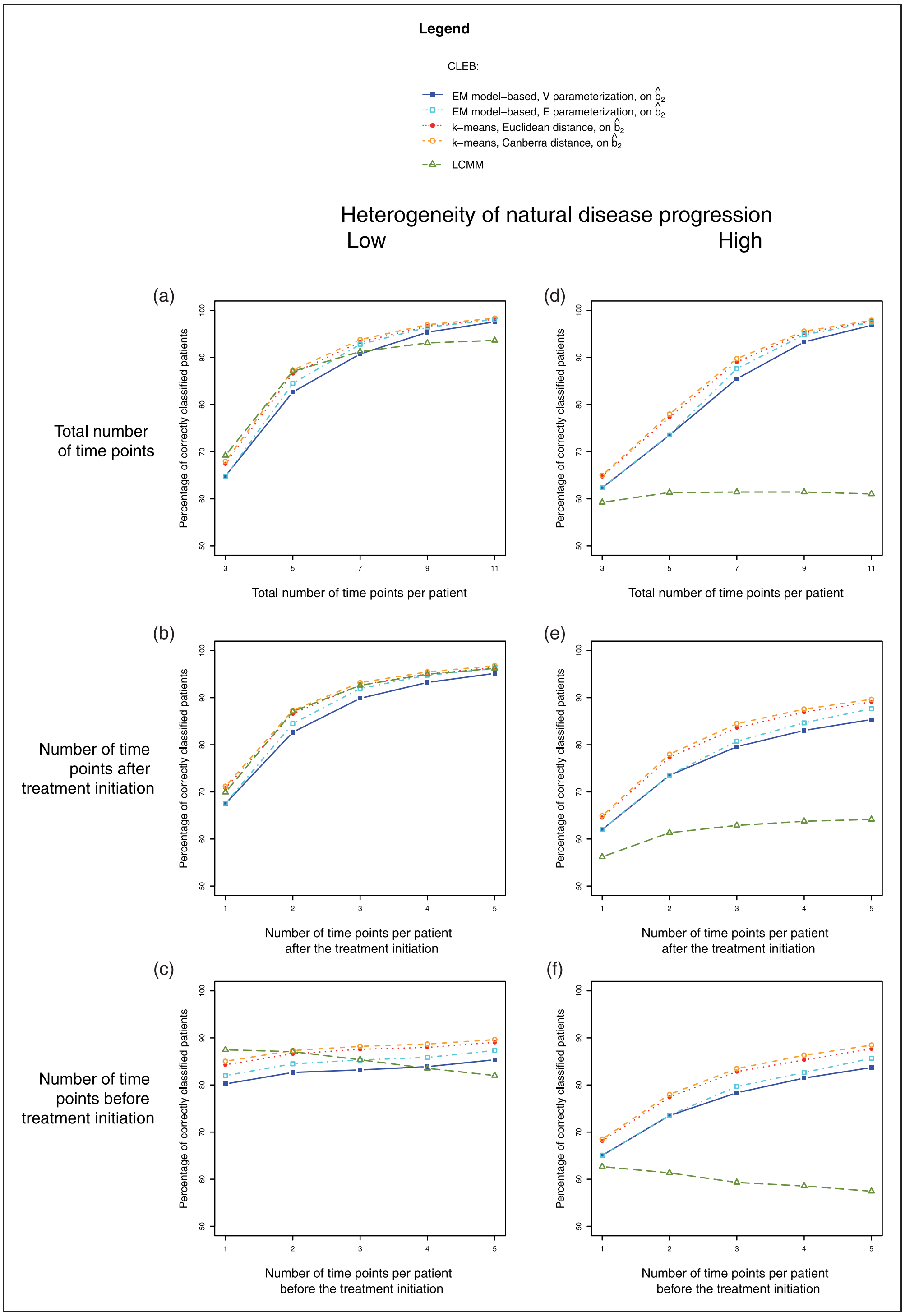


Figure 5. Impact of the number of time points on the percentage of correctly classified patients.

5 Applications

5.1 The impact of neuroleptics on the evolution of Huntington's disease

Huntington's disease is a rare and inherited neurodegenerative disorder caused by an expansion of a CAG (Cytosine-Adenine-Guanine) triplet repeat on the huntingtin gene on chromosome 4. It is characterized by choreiform movements, progressive dementia and psychiatric manifestations.[26] There is currently no cure and

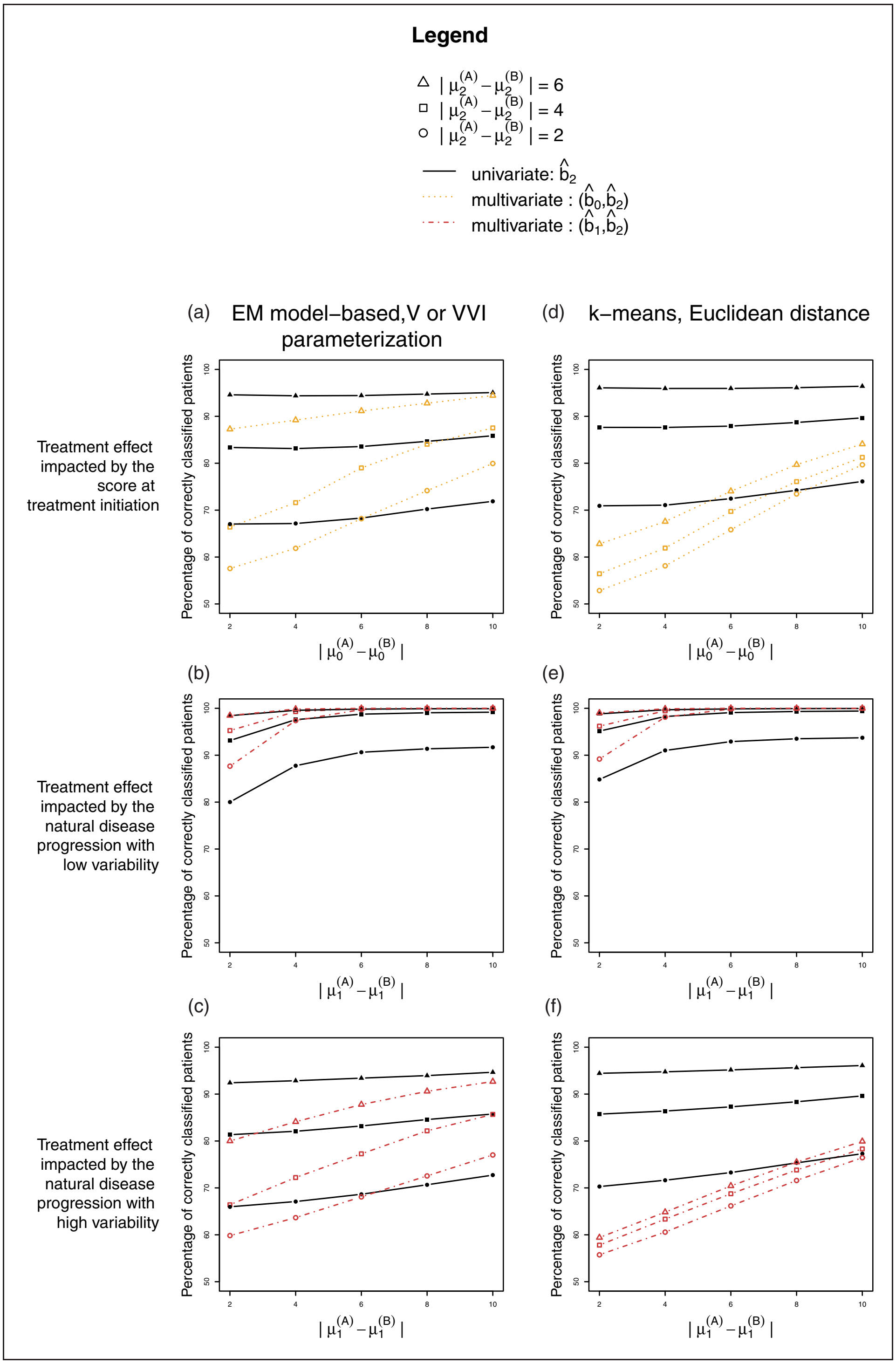


Figure 6. The case of a treatment effect on which baseline or evolution before treatment initiation has an impact.

all available treatments are symptomatic, i.e. they treat the symptoms but not the underlying disease. For example, AntiPsychotics and Related drugs (APRs) are commonly used for the treatment of chorea. Here, we evaluate the response to treatment with APRs.

We searched for responders to the treatment based on the evolution of the Functional Assessment Score (FAS), a clinical marker of the progress of the disease (score from 25 to 50). The treatment was supposed to start taking effect one month after the first prescription.

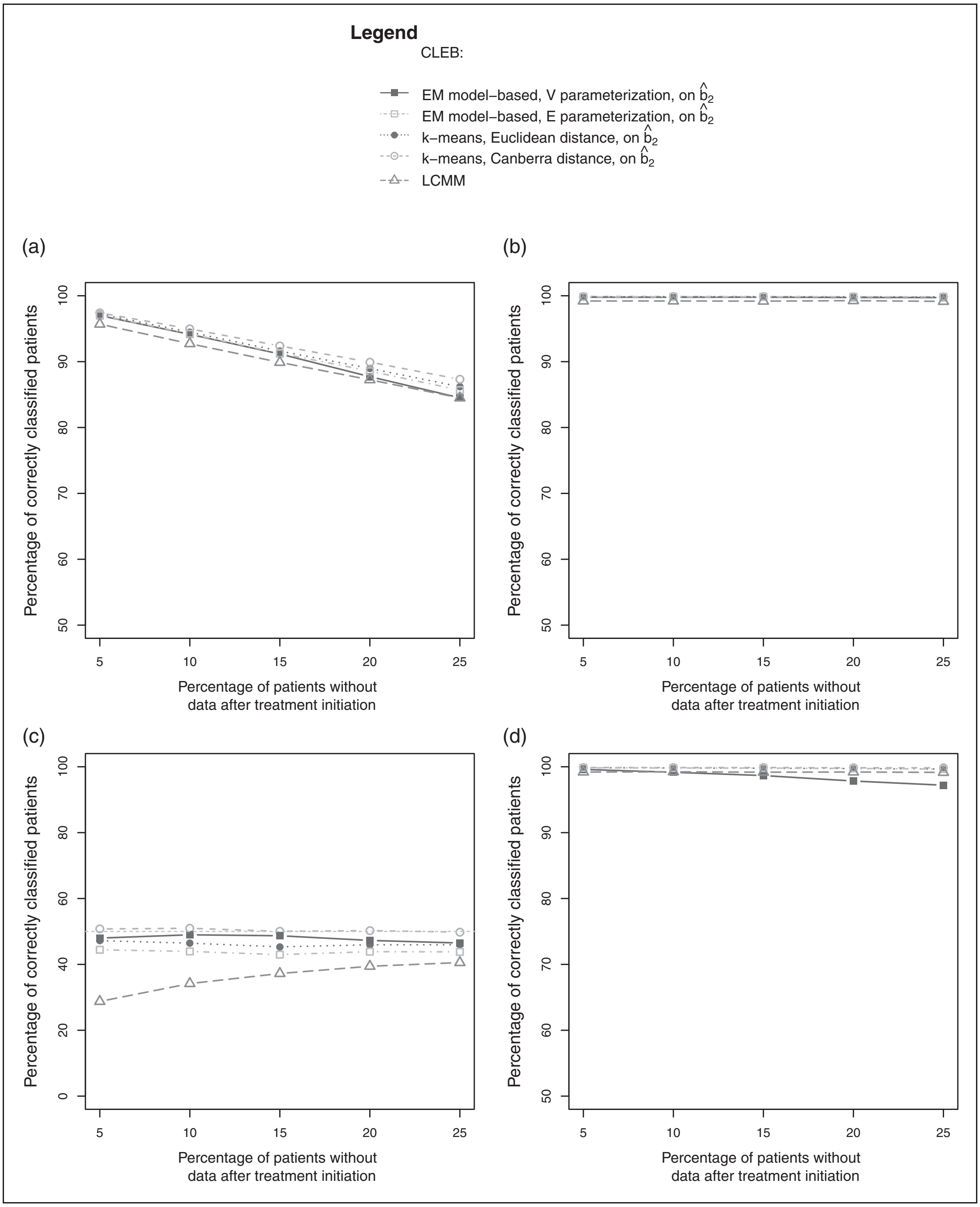


Figure 7. Impact of missing data on the classification. (a) Analysis with all patients (patients with and without missing data); (b) Analysis with complete cases (only patients without missing data); (c) Results for patients with missing data in analysis with all patients and (d) Results for patients without missing data in analysis with all patients.

Data were selected from the Huntington French Speaking Network cohort between 2002 and 2010, among the patients studied by Désaméricq et al.[27] In this clustering study, only 39 patients having APRs treatment who were followed at least twice before and twice after the treatment initiation were included. They were followed for 4.98 years ($SD = 1.58$), representing between 4 and 12 visits.

We applied the CLEB method (with the model-based algorithm and V parameterization) to the 39 patients. The CLEB splitted the population into two subgroups: 15 responders and 24 non-responders to APRs treatment. We then modelled the data with these groups of treatment responses as the covariate. The model showed an evolution

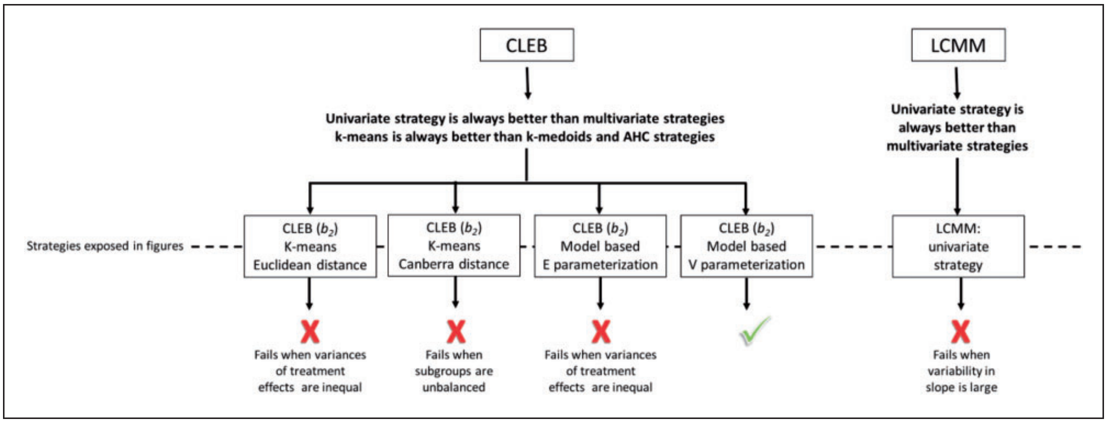

CLEB
Univariate strategy is always better than multivariate strategies
k-means is always better than k-medoids and AHC strategies
LCMM
Univariate strategy is always better than multivariate strategies
Strategies exposed in figures
CLEB (b_2) K-means Euclidean distance
CLEB (b_2) K-means Canberra distance
CLEB (b_2) Model based E parameterization
CLEB (b_2) Model based V parameterization
LCMM: univariate strategy
Fails when variances of treatment effects are inequal
Fails when subgroups are unbalanced
Fails when variances of treatment effects are inequal
Fails when variability in slope is large

Figure 8. Results of the simulation study.

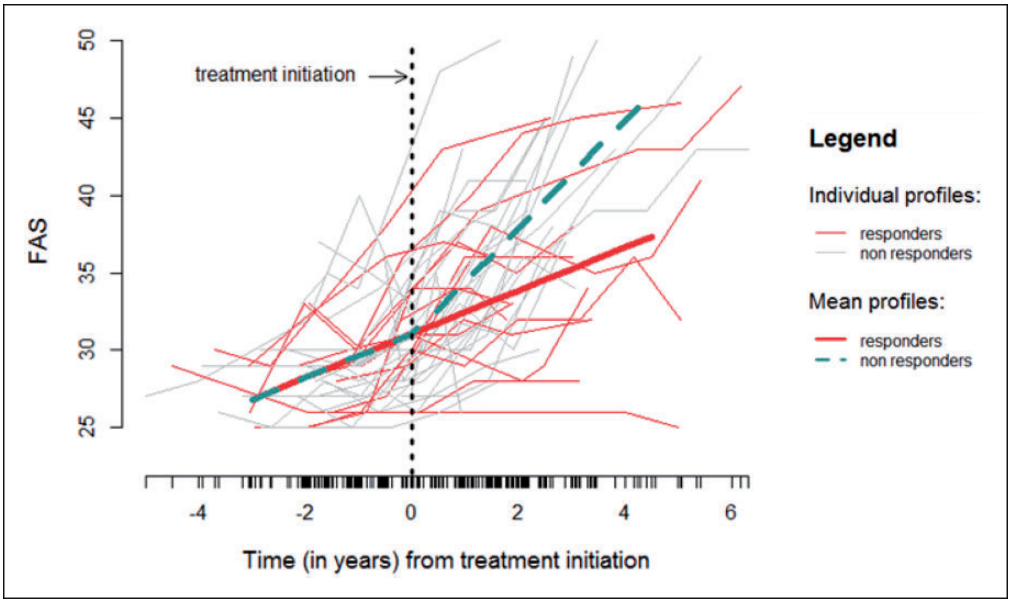


Figure 9. Evolution of FAS in the subgroups of responders and non-responders.

of FAS of 1.43 points per year ($SE = 0.22$, $P < 0.001$) during the baseline phase in the whole cohort. The difference in slope between the baseline phase and the treatment effect phase was –0.03 points per year ($SE = 0.35$, $P = 0.930$) for responders and 2.08 points per year ($SE = 0.37$, $P < 0.001$) for non-responders (Figure 9).

Patients with high CAG repeats are more frequently non-responders than those with low CAG repeats (Table 1). The repetition of CAG is correlated with the disease being more serious.[28–30] The clustering results suggest that APRs are inefficient for patients with high CAG repeats (the patients with the most severe symptoms). The two profiles of evolution we observed could reflect the treatment effect, the disease severity or both. The conclusion could only be speculative and a confirmatory analysis is required.

Repeating the method on the 741 subsets of the whole data set that were created by deleting two patients each time showed that the identified subgroups are robust, with only four patients being classified less than 90% of the time in the same subgroup.

The LCMM method was also applied to the data set. Sixty nine percentage of patients had a matching classification with both the CLEB and the LCMM methods (Table 2). The conclusions were similar with non-responders having higher CAG repeats (Mann-Whitney test, $P = 0.020$).

Table 1. Description of responders and non-responders to APRs.

	Whole cohort $N = 39$	Responders $N = 15$	Non-responders $N = 24$	*p*-values[b]
Age[a] (y)	50.07 (8.81)	52.31 (6.86)	48.67 (9.71)	0.260
Sex				
Male	23 (58.97%)	9 (60.00%)	14 (58.33%)	0.918
Female	16 (41.03%)	6 (40.00%)	10 (41.67%)	
Inheritance				
Paternal	19 (48.72%)	8 (53.33%)	11 (45.83%)	0.676
Maternal	17 (43.59%)	6 (40.00%)	11 (45.83%)	
Unknown	3 (7.69%)	1 (6.67%)	2 (8.33%)	
CAG	44.23 (3.06)	42.60 (2.47)	45.25 (3.00)	0.008
Age at onset (y)	43.54 (8.81)	46.64 (7.20)	41.65 (9.30)	0.106
Disease duration[a] (y)	7.38 (3.89)	6.86 (4.19)	7.70 (3.76)	0.387

[a] Measured at treatment initiation.

[b] Mann-Whitney test for quantitative data and chi square or Fisher exact test for qualitative data; y: in years; CAG: Cytosine-Adenine-Guanine; Responders and non-responders were defined by the CLEB algorithm.

Note: Quantitative data are expressed in mean (SD) and qualitative data in N(%).

Table 2. Concordance of responders and non-responders according to CLEB and LCMM methods.

	Responders LCMM	Non-responders LCMM
Responders CLEB	5	10
Non-responders CLEB	2	22

Note: CLEB: Clustering in Longitudinal data with Extended Baseline; LCMM: Latent-Class Mixed Model.

5.2 The impact of bariatric surgery on BMI

Obesity is an abnormal accumulation of body fat. It is associated with increased health problems, such as hypertension, Type II diabetes, coronary disease and hyperlipidemia. The Body Mass Index (BMI), obtained by dividing the weight by the square of the height, quantifies the tissue mass in an individual. Obesity is defined as a BMI score higher than 30. Currently, there are three categories of treatment: dietary modification, medication and surgery. Surgery treats people with potentially life-threatening obesity when other treatments, such as lifestyle changes, have not worked. Here, we evaluated two types of bariatric surgeries: sleeve gastrectomy and gastric bypass. Data were obtained from the records of a French bariatric centre. In the current clustering study, we analysed the period of 12 months before treatment initiation and 12 months after to assess the effect of the treatment on weight loss before stabilization. Only those 39 women with at least one measurement before surgery, one measurement at surgery and one measurement after surgery were included. They were followed for an average of 15.50 months (SD = 4.51), representing between 3 and 8 visits.

We applied the CLEB method (with the model-based algorithm and V parameterization) to the 39 women suffering from obesity. The CLEB split the population into two subgroups: 18 high-responders and 21 low-responders to surgery. We then modelled the data with these groups of treatment responses as the covariate. The model showed a stabilization of the BMI during the pre-operative period (mean of slope: –0.06 points per month, $SE = 0.06$, $P = 0.350$) in the whole cohort. The difference in the slope between the baseline phase and the treatment effect phase was –1.34 points per month ($SE = 0.11$, $P < 0.001$) for high-responders and –0.75 points per month ($SE = 0.10$, $P < 0.001$) for low-responders. Low-responders had a lower BMI at treatment initiation, with a BMI of of 8.58 points ($SE = 1.90$, $P < 0.001$) less than high-responders.

Younger women with a high weight at surgery are more frequently classified in the group of high-responders (Table 3). This is consistent with the fact that pre-operative BMI is positively associated with weight loss over a short follow-up period after bariatric surgery, whereas the correlation becomes negative over a longer follow-up period.[31] Moreover, younger patients might lose more weight because of their high metabolic activity compared to older patients.[32]

Table 3. Description of high-responders and low-responders to surgery.

	Whole cohort $N = 39$	High-responders $N = 18$	Low-responders $N = 21$	p-values[b]
Age[a] (y)	44.91 (10.11)	38.86 (7.91)	50.09 (8.93)	<0.001
Treatment				
Sleeve	19 (48.72%)	8 (44.44%)	11 (52.38%)	0.621
Bypass	20 (51.28%)	10 (55.56%)	10 (47.62%)	
Weight[a]	114.10 (20.76)	127.56 (21.67)	102.57 (10.80)	<0.001
Type 2 diabetes				
Yes	6 (16.67%)	3 (16.67%)	3 (16.67%)	>0.999
No	30 (83.33%)	15 (83.33%)	15 (83.33%)	
NA	3	0	3	
Sleep apnea				
Yes	7 (18.92%)	4 (22.22%)	3 (15.79%)	0.693
No	30 (81.08%)	14 (77.78%)	16 (84.21%)	
NA	2	0	2	
Hypertension				
Yes	9 (25.00%)	2 (11.11%)	7 (38.89%)	0.121
No	27 (75.00%)	16 (88.89%)	11 (61.11%)	
NA	3	0	3	

Note: Quantitative data are expressed in mean (SD) and qualitative data in N(%).

[a] Measured at treatment initiation.

[b] Mann-Whitney test for quantitative data and chi square or Fisher exact test for qualitative data; y: in years; NA: Not Available; High-responders and low-responders were defined by the CLEB algorithm.

BMI at surgery is linked to treatment effect, so we performed a multivariate clustering as a sensitivity analysis. We applied the CLEB method (with the model-based algorithm and VVI parameterization) on $(\hat{\boldsymbol{b}}_{\boldsymbol{0}}, \hat{\boldsymbol{b}}_{\boldsymbol{2}})$. The match between the univariate and the multivariate was 82% and the conclusions were similar, with a faster weight loss in younger women with higher weight and BMI at surgery initiation.

We also made a clustering that included 24 months post-surgery observations by changing the time into $log(\text{time} + 1)$ to avoid the non-linearity caused by the plateau in the stabilization period. Once again, the results match the previous analyses whether the clustering was univariate or multivariate.

6 Discussion

6.1 The CLEB algorithm with the EM model-based (V parameterization) strategy

In this paper, we have presented the CLEB method, a new method for classifying patients according to treatment efficacy in the case of continuous longitudinal data. The method has two steps. The first consists of modelling the entire trajectory of the data with measurements before and after treatment initiation. An extended baseline was used: data were modelled with two slopes, corresponding to the baseline and the treatment effect phases. The slope change appears at the assumed time of the treatment effect. Thus, the model has three mixed components: intercept $(\beta_0 + b_{0i})$, slope during baseline phase $(\beta_1 + b_{1i})$, and difference between the slopes at the baseline phase and the treatment effect phase $(\beta_2 + b_{2i})$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ is the vector of the fixed effects and $\boldsymbol{b_i} = (b_{0i}, b_{1i}, b_{2i})$ is the vector of the random effects for patient i. In the second step, the clustering is made on random predictions of $\boldsymbol{b_2}$ using the EM model-based algorithm assuming different variances (V parameterization) and allowing different shapes between clusters. The current study showed that the CLEB algorithm is useful for clustering patients according to treatment efficacy in the case of longitudinal data such as data obtained for a progressive disease (e.g. Huntington's disease[26]). The lag between treatment initiation and treatment effect has to be specified as an input parameter, but simulation studies showed that increase the variability of the lag does not have an impact on the results. This method is robust, regardless of the noise on the within- or between-subject variability of the baseline phase. Furthermore, the mixed model in the first step makes to the method insensitive to heterogeneity in the number and time of records between subjects. Even if the method could deal with missing data, patients need to have at least one measurement after treatment initiation to be included in the analysis.

However, there is a minimum number of time points required to make the method efficient, and there also must be time points before the treatment initiation if the natural disease progression is heterogeneous.

6.2 The lack of relevance of the other strategies

We considered other clustering strategies in the second step of the CLEB method. The simulation study showed that multivariate clustering (on $(\hat{b}_0, \hat{b}_2)$, $(\hat{b}_1, \hat{b}_2)$ or $(\hat{b}_0, \hat{b}_1, \hat{b}_2)$) could be preferred to univariate clustering (on $\hat{b}_2$) only if the treatment effect was linked to the patient's baseline conditions and the variance was low, whether, parametric or non-parametric strategies are used. However, when the variability is high, multivariate clustering strategies add more noise, and univariate strategies must be preferred.

For the non-parametric strategies, partitioning algorithms were more relevant than AHC algorithms, and the Canberra distance provided better results than the Euclidean one with two balanced subgroups but was inefficient for unbalanced subgroups. Indeed, in the case of univariate clustering, this distance will always separate positive and negative values.

In unbalanced scenarios, all the methods, except for the CLEB method with a model-based algorithm strategy, failed when $N_A = 2$. Even though this case seemed unrealistic, it was considered because of the possibility that a treatment had a beneficial effect only for a rare genetic profile.

6.3 Comparison of CLEB and LCMM

Furthermore, the CLEB method was compared to the LCMM method. For all the simulation scenarios, the CLEB method performed as well as or better than the LCMM method, especially when there was high variability of the slope before treatment initiation. Indeed, with the LCMM method, the definition of the subgroups and the estimation of the parameters were done simultaneously, leading to a greater influence of the baseline slope on the subgroup definition. For a large variance in treatment efficacy, both the CLEB and the LCMM methods became inefficient. Indeed, for large variance, the distribution of random terms, which is a mixture of K Gaussian distributions, tended to become unimodal.[33]

6.4 Some extensions for the CLEB method

For all the simulation studies, the number of clusters was an input parameter for the CLEB method. The choice of two clusters may make the model find two distinct subgroups even if they do not exist. However, the Bayesian Information Criterion may help in the choice of the number of clusters.

The CLEB method could have some extensions with more specific models. Indeed, we considered the case of a sustainable treatment effect, using a piecewise linear mixed model with two slopes. However, it was easy to extend this to the case of a piecewise linear mixed model with three slopes, the third corresponding to a plateau in treatment efficacy or to a resumption of the disease progression. Thus, subgroups of patients were defined according to short- and long-term treatment efficacy.

Furthermore, the assumption of a linear constant change for the outcome could be false. Indeed, the CLEB method proposed the use of a linear mixed model which assumed a constant change for the outcome. This assumption may not hold for psychometric scores characterized by upper and lower bounds. Considering the outcome as a discrete and bounded variable can improve the model and classify patients better. Indeed, it has been shown that, for handling this type of data, an alternative mixed model, handling this type of data, performed better than the classical linear mixed models[34] in data modelling. Splines or wavelets are also some modelling alternatives for specific outcomes such as time series data.[35]

Finally, only unsupervised algorithms, which attributed each patient to a cluster without a prior subgroup, were envisaged. However, if some patients could be easily identified as treatment responders or non-responders, with a mixture of labelled and unlabelled data, the algorithm would be improved by training it on the labelled patients and then applying it to the unlabelled patients as a semi-supervised algorithm.

6.5 Perspectives

This new method will help to define subgroups in the search for markers of treatment efficacy and to understand why some patients respond to treatment, while others fail to do so. It extracts information from pharmaco-epidemiological studies (the treatment arm of clinical trials or cohort studies with a treatment initiation during the follow-up). It is particularly interesting to find small subgroups of responders to a treatment that has never demonstrated its efficacy in a clinical trial. The definition of subgroups may help to find marker(s) of treatment response, which is a prerequisite for the implementation of stratified design for future clinical trials. This leads to therapies being matched with a specific patient population. It is anticipated that this will have a major effect on both clinical practice and the development of new drugs and diagnostics.[36]

Acknowledgements

The authors thank Désaméricq G., Dolbeau G., Verny C., Charles P., Durr A., Youssov K., Simonin C., Azulay J-P., Tranchant C., Goizet C., Damier P., Broussolle E., Demonet J-F., Morgado G., Cleret de Langavant L., Macquin-Mavier I. and Maison P. for their data of patients with Huntington's disease receiving antipsychotics and related drugs and Lazzati A. for his data of patients suffering from obesity and having a bariatric surgery. The authors thank Bardy P. from Centre des langues of Paris Descartes University and the proofreaders from Proof-Reading-Service.com for their english language correction.

The authors thank the neurologists and the neuropsychologists from the Multicentric Intracerebral Grafting in Huntington's Disease trial who collected the data for the simulation study: A-C. Bachoud-Lévi, MD, PhD (Henri Mondor hospital, Créteil, Principal investigator), M-F Boissé (Henri Mondor hospital, Créteil, Neuropsychologist), L. Lemoine (Henri Mondor hospital, Créteil, Neuropsychologist), C. Verny, MD, PhD (Angers hospital, Site coordinator), G. Aubin (Angers hospital, Neuropsychologist), J-F Demonet, MD, PhD (CHU Rangueil, Toulouse, Site coordinator), F. Calvas, MD (CHU Rangueil, Toulouse, Investigator), P. Krystkowiak, MD, PhD (Roger Salengro hospital, Lille and and CHU d'Amiens, Amiens, Sites coordinator), C. Simonin, MD, PhD (Roger Salengro hospital, Lille, Investigator), M. Delliaux (Roger Salengro hospital, Lille, Neuropsychologist), P. Damier, MD, PhD (Nord Laennec hospital, Nantes, Site coordinator), P. Renou (Nord Laennec hospital, Nantes, Investigator), F. Supiot (Erasme hospital, Bruxelles, Site coordinator), H. Slama (Erasme hospital, Bruxelles, Neuropsychologist).

Declaration of conflicting interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Mrs. Schramm was successively supported by the NeuroStemcell Consortium (European Community Seventh Framework Program grant Agreement No. 222943) and by "Investments for the future" (ANR-11-INBS-0011 - NeurATRIS: Infrastructure de recherche translationnelle pour les biothrapies en Neurosciences).

Dr. Katsahian reports no disclosures.

Dr. Vial reports no disclosures.

Dr. Bachoud-Lévi was consultant for Teva once in 2014. She received grants from the Ministry of Health supporting the National reference center for Huntington's disease and several grants for academic trials provided by the "Direction de la Recherche Clinique" (AP-HP). She is partner from several investments for the future (Labex IEC, Neuratris) and a FP7 EU grant FP7 (Repair-HD).

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by investment for the future NeurATRIS: Infrastructure de recherche translationnelle pour les biothrapies en Neurosciences (ANR-11-INBS-0011), European Community Seventh Framework Program Neurostemcell (Grant Agreement no. 222943), European Community Seventh Framework Program Repair-HD (Grant Agreement no 602245). The Département d'Études Cognitives of the École Normale Supérieure is supported by two ANR grants from the French Research Agency (ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL).

References

1. Everitt BS, Landau S, Leese M, et al. *Cluster analysis*, 5th ed. Chichester, West Sussex, U.K: Wiley-Blackwell, 2011.
2. Koestler DC, Marsit CJ, Christensen BC, et al. A recursively partitioned mixture model for clustering time-course gene expression data. *Trans Cancer Res* 2014; **3**: 217.
3. Harrington M, Velicer WF and Ramsey S. Typology of alcohol users based on longitudinal patterns of drinking. *Addict Behav* 2014; **39**: 607–621.
4. Castellini G, Fioravanti G, Sauro CL, et al. Latent profile and latent transition analyses of eating disorder phenotypes in a clinical sample: a 6-year follow-up study. *Psych Res* 2013; **207**: 92–99.

5. Kent P and Kongsted A. Identifying clinical course patterns in sms data using cluster analysis. *Chiropract Manual Therap* 2012; **20**: 1–12.
6. Tepper PG, Randolph JF Jr, McConnell DS, et al. Trajectory clustering of estradiol and follicle-stimulating hormone during the menopausal transition among women in the study of women's health across the nation (swan). *J Clin Endocrinol Metabol* 2012; **97**: 2872–2880.
7. Muthén B and Muthén LK. Integrating person-centered and variable-centered analyses: growth mixture modeling with latent trajectory classes. *Alcoholism: Clin Exp Res* 2000; **24**: 882–891.
8. Verbeke G and Lesaffre E. A linear mixed-effects model with heterogeneity in the random-effects population. *J Am Stat Assoc* 1996; **91**: 217–221.
9. Genolini C and Falissard B. Kml: k-means for longitudinal data. *Comput Stat* 2010; **25**: 317–328.
10. Madsen K, Miller J and Province M. The use of an extended baseline period in the evaluation of treatment in a longitudinal duchenne muscular dystrophy trial. *Stat Med* 1986; **5**: 231–241.
11. Laird NM and Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**: 963–974.
12. Laird NM. Missing data in longitudinal studies. *Stat Med* 1988; **7**: 305–315.
13. Fraley C and Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc* 2002; **97**: 611–631.
14. Redner RA and Walker HF. Mixture densities, maximum likelihood and the em algorithm. *SIAM Rev* 1984; **26**: 195–239.
15. Banfield JD and Raftery AE. Model-based gaussian and non-gaussian clustering. *Biometrics* 1993; **49**: 803–821.
16. Hartigan JA and Wong MA. Algorithm as 136: a k-means clustering algorithm. *Appl Stat* 1979; **28**: 100–108.
17. MacQueen J, et al. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Oakland, CA, USA, vol. 1, pp. 281–297.
18. Kaufman L and Rousseeuw P. Statistical data analysis based on the Ll norm. In: dodge Y (ed.) *Clustering by means of medoids*. Amsterdam: North-Holland, 1987, pp. 405–416.
19. Day WH and Edelsbrunner H. Efficient algorithms for agglomerative hierarchical clustering methods. *J Class* 1984; **1**: 7–24.
20. Team RDC. R: a language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria, http://www.R-project.org (2013).
21. Pinheiro J, Bates D, DebRoy S, et al. nlme: linear and nonlinear mixed effects models. *R package version 3.1-105*. 2012, http://CRAN.R-project.org/package = nlme
22. Fraley C, Raftery AE, Murphy TB, et al. mclust: normal mixture modeling for model-based clustering, classification, and density estimation. *R package version 4.3*. 2012, http://CRAN.R-project.org/package=mclust
23. Lucas A. *amap: another multidimensional analysis package*, http://CRAN.R-project.org/package=amap. R package version 0.8-12 (2014).
24. Maechler M and Rousseeuw P. *cluster: cluster analysis basics and extensions. R package version 1.14.3*. 2012, http://CRAN.R-project.org/package=cluster
25. Proust-Lima C, Philipps V, Diakite A, et al. *lcmm: estimation of extended mixed models using latent classes and latent processes*, http://CRAN.R-project.org/package=lcmm. R package version 1.6.6 (2014).
26. Bates G, Harper P and Jones L. *Huntington's disease: oxford monographs on medical genetics*. New York: Oxford University Press, 2002.
27. Désaméricq G, Dolbeau G, Verny C, et al. Effectiveness of anti-psychotics and related drugs in the huntington french-speaking group cohort. *PLoS ONE* 2014; **9**: e85430. DOI:10.1371/journal.pone.0085430
28. Trottier Y, Biancalana V and Mandel JL. Instability of cag repeats in huntington's disease: relation to parental transmission and age of onset. *J Med Gen* 1994; **31**: 377–382.
29. Brandt J, Bylsma F, Gross R, et al. Trinucleotide repeat length and clinical progression in huntington's disease. *Neurology* 1996; **46**: 527–531.
30. Langbehn DR, Hayden MR and Paulsen JS. Cag-repeat length and the age of onset in huntington disease (hd): a review and validation study of statistical approaches. *Am J Med Gen Part B: Neuropsych Gen* 2010; **153**: 397–408.
31. Livhits M, Mercado C, Yermilov I, et al. Preoperative predictors of weight loss following bariatric surgery: systematic review. *Obes Surg* 2012; **22**: 70–89.
32. Agüera Z, García-Ruiz-de Gordejuela A, Vilarrasa N, et al. Psychological and personality predictors of weight loss and comorbid metabolic changes after bariatric surgery. *European Eating Disorders Rev* 2015; **23**: 509–516.
33. Strenio JF, Weisberg HI and Bryk AS. Empirical bayes estimation of individual growth-curve parameters and their relationship to covariates. *Biometrics* 1983; **39**: 71–86.
34. Proust-Lima C, Dartigues JF and Jacqmin-Gadda H. Misuse of the linear mixed model when evaluating risk factors of cognitive decline. *Am J Epidemiol* 2011; **174**: 1077–1088.
35. James GM and Sugar CA. Clustering for sparsely sampled functional data. *J Am Stat Assoc* 2003; **98**: 397–408.
36. Trusheim MR, Burgess B, Hu SX, et al. Quantifying factors for the success of stratified medicine. *Nature Rev Drug Discov* 2011; **10**: 817–833.