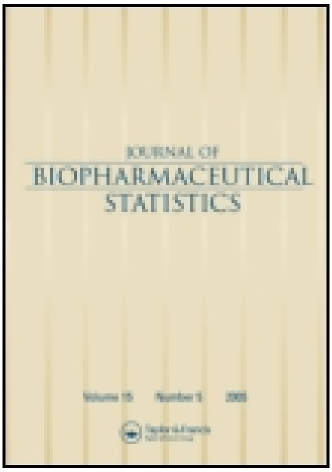




Journal of Biopharmaceutical Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lbps20

# On Comparison of Clustering Methods for Pharmacoepidemiological Data

Fanny Feuillet[ab], Lise Bellanger[c], Jean-Benoit Hardouin[ab], Caroline Victorri-Vigneau[ad] & Véronique Sébille[ab]

[a] EA 4275 SPHERE Biostatistics, Pharmacoepidemiology and Human Sciences Research, University of Nantes, Nantes, France

[b] Plateforme de Biométrie, Cellule de promotion de la recherche clinique, Nantes University Hospital, Nantes, France

[c] UMR CNRS 6629, University of Nantes, Nantes, France

[d] Centre for Evaluation and Information on Pharmacodependence, Nantes University Hospital, Nantes, France

Accepted author version posted online: 06 Jun 2014.

To cite this article: Fanny Feuillet, Lise Bellanger, Jean-Benoit Hardouin, Caroline Victorri-Vigneau & Véronique Sébille (2015) On Comparison of Clustering Methods for Pharmacoepidemiological Data, Journal of Biopharmaceutical Statistics, 25:4, 843-856, DOI: 10.1080/10543406.2014.920855

To link to this article: http://dx.doi.org/10.1080/10543406.2014.920855

*Journal of Biopharmaceutical Statistics*, 25: 843–856, 2015

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543406.2014.920855

Taylor & Francis
Taylor & Francis Group



# ON COMPARISON OF CLUSTERING METHODS FOR PHARMACOEPIDEMIOLOGICAL DATA

**Fanny Feuillet[1,2], Lise Bellanger[3], Jean-Benoit Hardouin[1,2], Caroline Victorri-Vigneau[1,4], and Véronique Sébille[1,2]**

[1]*EA 4275 SPHERE Biostatistics, Pharmacoepidemiology and Human Sciences Research, University of Nantes, Nantes, France*
[2]*Plateforme de Biométrie, Cellule de promotion de la recherche clinique, Nantes University Hospital, Nantes, France*
[3]*UMR CNRS 6629, University of Nantes, Nantes, France*
[4]*Centre for Evaluation and Information on Pharmacodependence, Nantes University Hospital, Nantes, France*

***The high consumption of psychotropic drugs is a public health problem. Rigorous statistical methods are needed to identify consumption characteristics in post-marketing phase. Agglomerative hierarchical clustering (AHC) and latent class analysis (LCA) can both provide clusters of subjects with similar characteristics. The objective of this study was to compare these two methods in pharmacoepidemiology, on several criteria: number of clusters, concordance, interpretation, and stability over time. From a dataset on bromazepam consumption, the two methods present a good concordance. AHC is a very stable method and it provides homogeneous classes. LCA is an inferential approach and seems to allow identifying more accurately extreme deviant behavior.***

***Key Words:*** Agglomerative hierarchical clustering; Clusters comparison; Clusters stability; Drug dependence; Multiple correspondence analysis; Latent class analysis.

## 1. INTRODUCTION

France has one of the highest recorded rates of psychotropic drug use (anxiolytics, hypnotics, antidepressants) compared to other countries and this constitutes a national public health problem. Initially limited to the most severe disorders, their use has now been extended to less serious disorders and has gradually become commonplace. According to epidemiological surveys, more than one person over three has already used a psychotropic drug during his life (Gasquet et al., 2005; Briot, 2006). Particularly, benzodiazepines were one of the most prescribed psychotropic drugs worldwide (bromazepam for example). This class of psychotropic drugs is indicated for the treatment of anxiety disorders or sleeps disorders. However, many studies reported their abuse and dependence potential (Cadet-Taïrou et al., 2008). Overconsumption can have serious health, social and economic consequences.

Received October 18, 2012; Accepted February 7, 2014
Address correspondence to Fanny Feuillet, EA 4275 SPHERE Biostatistics, Pharmacoepidemiology and Human Sciences Research, University of Nantes, 1 rue Gaston Veil BP 53508, 44035 Nantes Cedex 1, France; E-mail: fanny.feuillet@univ-nantes.fr

Indeed, various studies suggest the existence of a link between consumption of benzodiazepines and dementia (Verdoux et al., 2005; Billioti de Gage et al., 2012). Other risks are highlighted as memory disorders or withdrawal syndrome. There is also a problematic use of benzodiazepines with abuse among drug users (Cavalie and Richard, 2012). Therefore, monitoring of misuse, abuse, and dependence is recommended. To assess these risks, the Centre for Evaluation and Information on Pharmacodependence (CEIP) has been commissioned in France. The three main missions of the CEIP are: (i) to collect data and assess the potential of dependence on psychoactive drugs, (ii) to provide information on the risk of abuse or dependence of psychoactive substances, and (iii) to carry out research.

For pharmacodependence assessment, health authorities insist on the importance of obtaining quantitative information on dose and co-prescriptions of consumers, to identify their consumption characteristics (Briot, 2006). In this context of post-marketing surveillance of psychotropic drugs use, it has been emphasized that it would be interesting to use statistical methods to identify profiles of psychotropic drugs users. Health authorities rely on these pharmacoepidemiological studies to quantify drug dependence and to implement effective measures to reduce psychotropic drug consumption and associated risks.

The use of data from the General Health Insurance Scheme (GHIS) is recommended by health authorities. The GHIS covers 80% of the French population. Our database however excludes farmers and independent professions (about 20%). These large databases contain all reimbursement of dispensing for a given drug.

Identification of different consumption behavior could help to characterize the consumption of psychotropic drugs, to better target existing problems. Several clustering methods are available to make clusters of subjects that have similar characteristics.

Latent class analysis (LCA) was proposed by Lazarsfeld (1950) as an inferential statistical method, used to identify groups of individuals (McCutcheon, 1987). It postulates the existence of a latent variable, not directly observable but whose effects can be observed. This method is based on the assumption of local independence, meaning that manifest observed variables are associated because the population is a mixture of two or more classes. Accordingly, the associations between these variables should be explained by the class membership. A growing number of epidemiological studies have applied this method in order to determine subtypes of users of different psychoactive drugs (Chung et al., 2006; Reboussin et al., 2006; Agrawal et al., 2007; Kendler et al., 2013).

Agglomerative hierarchical clustering (AHC) is one of the leading data descriptive methods. It is aimed at grouping individuals with similar pattern of responses from quantitative data. The objective is to class individuals into groups that are as homogeneous as possible (Anderberg, 1973). AHC was applied to group clonazepam users, based on data from health reimbursement system. This study allowed differentiating and describing two subgroups of consumers: subjects with or without a deviant behavior (Frauger et al., 2006, 2003). AHC was also applied in a slightly different context (use of statins) (Latry et al., 2010) but with the same objective to characterize a drug consumption.

Application of these methods in a context of post-marketing phase is very important. Indeed, pharmacodependence cannot be assessed in clinical trials, because it is a rare side effect and patients at risk of addiction are excluded from these trials.

Can we identify a method, between LCA or AHC, or a combination of both, that would be most appropriate to characterize drug consumption behavior? What are the most interesting properties of each method regarding this objective? The aim of this study is to compare the application of these two clustering methods, commonly used in a pharmacodependence context. Each method was performed to identify and characterize groups of users with similar consuming behaviors, using available variables describing subject's bromazepam consumption. In this paper, we compare and analyze properties of each method. Comparison of obtained partitions (number of clusters, pharmacoepidemiological interpretation) and stability over time are used as evaluation criteria.

## 2. MATERIAL AND METHODS

### 2.1. Material

**2.1.1. Data.** In France, all benefits reimbursed to patients affiliated to the GHIS are registered to make up an extensive database of all drugs dispensing in France. From the population affiliated to the GHIS in the Pays de Loire region (more than 3.5 million inhabitants), we had selected the subjects who were aged over 18 years and who had received at least two dispensing of bromazepam at two different times during the first half of 2008. Two dispensing were needed to calculate an average daily consumption. The same criteria were used on the first half of 2009.

**2.1.2. Studied Drug.** The drug bromazepam belongs to the benzodiazepines class and is mainly marketed under the name of Lexomil®. It is recommended for symptomatic treatment of severe and/or incapacitating anxiety manifestations and for prevention and treatment of delirium tremens and other signs of alcohol withdrawal. In the literature, it has been observed that a significant proportion of users display misuse behavior, looking for positive effects. These effects correspond to removed inhibitions, often in combination with other substances such as alcohol or cannabis (Cadet-Taïrou et al., 2008).

**2.1.3. Assessment and Measures.** Each clustering method (AHC and LCA) was performed on six available binary variables describing subject's bromazepam consumption behavior, defined by pharmacological experts (Wainstein et al., 2011):

- Overconsumption: a consumption factor (CF) was defined as the estimated average daily consumption divided by the maximum recommended daily dose. Average daily consumption (in mg/days) is equal to the delivered quantity divided by the length of treatment. The threshold for overconsumption was defined by a value of the CF greater than 1.
- General practitioner prescription: psychotropics drugs could be prescribed by a general practitioner or by a specialist. If at least one prescription was made by a specialist then a specialist prescription was considered.
- Doctor shopping behavior: patients may develop a behavior known as "doctor shopping," which means that the same patient obtained several dispensing of bromazepam prescribed by different physicians in the same time frame. The threshold for "doctor shopping" was defined by more than 3 prescribing physicians during the 6-months study period.
- Pharmacy shopping behavior: patients may also develop a behavior known as "pharmacy shopping," which means that the same patient consults several pharmacies who deliver bromazepam in the same time frame. The threshold for "pharmacy shopping" was defined by more than 3 pharmacies who delivered the drug during the 6-months study period.

French practice guidelines on the suitable use of psychotropic drugs help physicians to prescribe them as well as possible associations. From these guidelines, two binary variables have been created:

- Prescription in agreement with practice guidelines relating to the therapeutic class of the studied psychotropic drug. A patient was not in compliance with the guidelines if he(she) had received one or more other benzodiazepines than bromazepam during the study period.
- Prescription in agreement with practice guidelines relating to other classes of associated psychotropic drugs. A patient was not in compliance with the guidelines if he(she) had received two delivering of antidepressant drugs and/or two delivering of antipsychotic drugs and/or two delivering of maintenance treatment (methadone or buprenorphine) during the study period.

These binary variables were used to define consumption profiles.



## 2.2. Methods

**2.2.1. Latent Class Analysis.** LCA is a statistical method based on a formal model. It is a particular mixture model, suitable for categorical data. It aims at identifying subgroups of individuals from categorical data. It postulates the existence of a latent variable, not directly observable but whose effects can be observed, such as consumption behavior (McCutcheon, 1987).

A latent class model is defined by $p$ manifest observed binary variables $\left(X^1, X^2, \ldots, X^p\right)$, $Y$ the latent variable with $C$ classes and $\boldsymbol{x_i} = (x_i^1, \ldots, x_i^p)$ a vector of binary responses for individual $i (i = 1, \ldots, n)$. We consider $x_i$ as the response profile for individual $i$ (indicators of consumption behavior through manifest observed variables).

LCA is characterized by two sets of parameters (McCutcheon 1987; Hagenaars and McCutcheon, 2002):

1. Class probability: $\pi_m = Pr(Y = m)$, is the probability of being in latent class $m$ $(m = 1, \ldots, C)$
2. Conditional probability: $p_{jm} = Pr\left(X^j = 1 | Y = m\right)$ with $j = 1, \ldots, p$, is the probability that an individual in latent class $m$ is involved with behavior $X^j$

This method is based on the assumption of local independence, that is, conditional on latent class membership, the manifest variables are mutually independent of each other. As a consequence, it is hypothesized that the associations between consumption behaviors (manifest variables) should only be explained by the class membership (individuals belong to different latent classes) (McCutcheon, 1987; Magidson and Vermunt, 2004).

Parameters are estimated by the method of maximum likelihood with an expectation maximization algorithm. The log likelihood is defined as

$$l(\pi_m, p_{jm} | x_i) = \sum_{i=1}^{n} ln \left( \sum_{m=1}^{C} \pi_m \prod_{j=1}^{p} p_{jm}^{x_i^j} (1 - p_{jm})^{1 - x_i^j} \right)$$

The first aim of LCA is to determine the smallest number of latent class $C$ that could explain the relationships observed among the manifest variables. Model selection was applied to statistically assess the fit of different latent class models with 2, 3, . . . up to the maximum plausible number of latent classes, defined to be 8. This choice is defined taking into account pharmacological knowledge of an expert in the field, possibilities of interpretation and also statistical consideration (the small number of available variables to define classes). To take into account problems related to local optima, we repeated this procedure by using 100 random starting values and selected the most parsimonious model that had an acceptable fit to the observed data (Magidson and Vermunt, 2004). Bayesian information criterion (BIC) was useful to comparing and selecting models. Model with the lowest BIC value was selected (Nylund et al., 2007).

In case of local dependence, manifest variables were assumed to be associated within latent classes. To identify this situation we used a diagnostic index called bivariate residual (BVR), that corresponds to Pearson chi-square statistic divided by the degrees of freedom. A BVR value larger than 1 suggested existence of local dependence between the two variables studied (Magidson and Vermunt, 2004). To take into account detected local dependencies, we included direct effects parameters. To do so, we combined each pair of dichotomous variables (responsible for the local dependence) into one item, with four response categories. This new item was included in the latent class model (Owen and Videras, 2006).

The final step was to classify subjects into appropriate latent classes. Information about an individual's class membership can be expressed through a set of posterior probabilities. We obtained the posterior probability that an individual $i$ with $x_i$ response vector belongs to latent class $m$. An individual was assigned in the class where he(she) had the highest estimated posterior probability (modal assignment).

$$\widehat{Pr}\left(Y = m | X_i = \boldsymbol{x_i}\right) = \frac{\hat{\pi}_m \prod_{j=1}^{p} \hat{p}_{jm} * x_i^j (1 - \hat{p}_{jm})^{1-x_i^j}}{\sum_{m=1}^{C} \hat{\pi}_m \prod_{j=1}^{p} \hat{p}_{jm} * x_i^j \left(1 - \hat{p}_{jm}\right)^{1-x_i^j}}$$

The software Latent Gold, version 4.5, was used for latent class model fitting.

**2.2.2. Multiple Correspondence Analysis and Agglomerative Hierarchical Clustering.** The second method used to identify structure in the data was agglomerative hierarchical clustering (AHC), which provides clusters from a given dissimilarity matrix (Benzécri, 1973). But, the original variables describing user's consumption behavior were coded as binary variable. It was necessary in a primary step to use multiple correspondence analysis (MCA) to produce factor projections. MCA is very popular in the French literature and has reached a high level of development and use (Lebart et al., 2000). It appears to be the counterpart of principal component analysis (PCA) for categorical data, used to detect and represent underlying structures in a dataset as points in a low-dimensional space.

MCA is an extension of correspondence analysis applied to the whole dataset, coded as binary variables called indicator matrix. Suppose that there are $p$ variables, each variable $j$ having $J_j$ categories, with $J = \sum_{j=1}^{p} J_j$ denoting the total number of categories. The indicator matrix, denoted $\mathbf{Z}$ is composed of $J$ columns. The total inertia of the indicator matrix is

$$Inertia\left(\mathbf{Z}\right) = \sum_{j=1}^{p} \left(\frac{J_j - 1}{p}\right) = \frac{J - p}{p} = \frac{J}{p} - 1$$

Thus, the total inertia depends only on the average number of modalities per variables. In our case, the variables have all the same number of modalities ($\forall j$, $J_j = 2$) and therefore $Inertia\,(\mathbf{Z}) = 1$.

Since $J - p$ is the dimensionality of $\mathbf{Z}$, the average inertia per dimension, is $\frac{1}{p}$. In practice, the value $\frac{1}{p}$ serves as a threshold for deciding which axes are worth interpreting (analogous to the threshold of 1 for the eigen-values in PCA). The values in the remaining dimensions, therefore, tend to be small and may be dropped with minimal loss of information. As PCA, MCA could be used for dimensionality reduction or data table approximation in a lower dimension. With all axes, we reconstructed exactly the data matrix. If we dropped a part of the axes, we partially reconstructed it and implicitly assumed that the neglected part was "noise" similar to a random part of a model of reconstitution of the matrix data. To approximate the data table, we decided to retain a number of axes leading to at least 60% of the total inertia, even if some retained axes generate a proportion of inertia less than $\frac{1}{p}$.

In a second step, AHC with Ward's method (Ward, 1963), was applied on the data table approximation represented by the first MCA axes. Hierarchical clustering is a nested set of partitions represented by a dendrogram. The agglomerative algorithm begins with $n$ subclusters, each containing a single data point, and at each stage merges the two most similar groups to form a new cluster. The algorithm proceeds until all the data are in a single cluster. Ward's method is appropriate for the clustering of points in Euclidean space (Lebart et al., 2000). Its aim is to minimize the total within-group sum of squares. The optimal number of clusters was determined visually by the best cut on the dendrogram: large changes in fusion levels were taken to indicate the best cut.

MCA and AHC were performed using SAS 9.2 software.

**2.2.3. Comparison of Partitions and Stability.** The $n$ consumers of bromazepam were divided into $k$ different response profiles ($k \leq 2^6$), defined from the $p$ binary available variables previously described ($p = 6$). Therefore, each response profile contained a different number of consumers.

The two methods (LCA and AHC) were compared according to several criteria: number of clusters, concordance between clusters coming from both methods and stability over time. To select the number of clusters, BIC criterion was used for LCA, and mainly by visual analysis of the dendrogram for AHC.

To measure concordance between clusters, confusion matrix was constructed between partitions obtained with the two methods. Confusion matrix corresponds to a two-way frequency table with LCA partition and AHC partition. Then concordance between the two partitions was calculated using coefficients measures:

a. Kappa coefficient is a measure of concordance between two qualitative partitions, taking into account the proportion of agreement due to chance (Cohen, 1960).
b. Jaccard and Rand index measures the percentage of agreement between the two partitions (Hubert and Arabie, 1985). Theses coefficients have a value between 0 and 1, with 0 indicating that the two data clusters have a poor agreement and a value of 1 implies perfect agreement.

Coefficients of concordance were calculated on the number of users. For confusion matrix, results are also presented as response profiles in order to facilitate interpretation.

A pharmacological stability was expected between 2008 and 2009, since prescription guidelines were not modified during this time frame. To evaluate this stability over time, each clustering method was performed in 2008 and in 2009. A confusion matrix was constructed including LCA clusters in 2009 and LCA clusters in 2008. Kappa coefficient was used to measure concordance between the two years. Since the number of consumers was different between the two years, this coefficient was calculated from the number of response profiles. The same reasoning was used for AHC.

To assess homogeneity of each partition, intra-class inertia was used. This overall measure of classes' homogeneity corresponds to average inertia of the $C$ classes, weighted by the corresponding size.

## 3. RESULTS

### 3.1. Description

These samples include data from 40,644 bromazepam users in 2008 and 44,756 in 2009, which represents 53 responses profiles each year. With the 6 dichotomous observed variables, that described consumption behaviors, there were $2^6 = 64$ possible responses profiles; so, 11 were not represented in our samples. Table 1 shows characteristics of these users for the two years. In 2009, among these users, about 74% were women and average age was 62 years. More than four out of five times, the prescription was made by a general practitioner. Overconsumption proportion was low (1.2% of users had a CF greater than 1), as were doctor shopping (0.4%) and pharmacy shopping (1.3%). Nearly 40% of prescriptions were not in agreement with practices guidelines related to the therapeutic class of bromazepam and 6% were not in agreement with practices guidelines related to other classes of associated psychotropic drugs. In 2008, we observed similar proportions for all observed variables. Prevalence of consumption behaviors appears to be stable between the two years.

### 3.2. Latent Class Analysis

Using LCA, the best-fitting model was the 4-class model. For both years, local dependencies were detected for four pairs of items that were included as direct effects: "kind of prescribing physicians" with "doctor shopping," "prescription in agreement with practice guidelines relating to the therapeutic class" with "consumption factor" and with "doctor shopping," "prescription in agreement with practice guidelines relating to other classes,"

**Table 1** Description of characteristics and consumption behaviors among bromazepam users in 2008 and 2009

| | 2008 | 2009 |
|---|---|---|
| | $n = 40644$ | $n = 44756$ |
| Age (mean ± std) | 62.0 ± 15.2 | 62.4 ± 15.2 |
| Female | 73.7% | 73.6% |
| Overconsumption | 1.1% | 1.2% |
| « Doctor shopping » | 0.4% | 0.4% |
| « Pharmacy shopping » | 1.2% | 1.3% |
| General practitioner prescription | 85.9% | 86.0% |
| Not in agreement to the therapeutic class | 39.2% | 38.1% |
| Not in agreement to other classes | 7.2% | 6.1% |

**Table 2** Description of latent class models after modal assignment – 2008 and 2009

| | 2008 | | | | 2009 | | | |
|---|---|---|---|---|---|---|---|---|
| | Class 1 | Class 2 | Class 3 | Class 4 | Class 1 | Class 2 | Class 3 | Class 4 |
| *Class repartition (%)* | 58.0 | 33.1 | 8.6 | 0.4 | 61.0 | 30.6 | 8.1 | 0.4 |
| Overconsumption (%) | 0.0 | 1.4 | 4.4 | 76.6 | 0.2 | 1.3 | 4.0 | 79.8 |
| « Doctor shopping » (%) | 0.0 | 0.0 | 1.8 | 55.8 | 0.0 | 0.0 | 2.4 | 51.4 |
| « Pharmacy shopping » (%) | 0.5 | 0.7 | 3.6 | 96.8 | 0.5 | 0.9 | 3.9 | 95.6 |
| General practitioner prescription (%) | 90.6 | 100.0 | 1.7 | 51.3 | 90.3 | 100.0 | 1.2 | 57.9 |
| Not in agreement to the therapeutic class (%) | 0.0 | 95.0 | 88.0 | 73.4 | 0.0 | 100.0 | 88.7 | 79.8 |
| Not in agreement to other classes (%) | 0.0 | 12.8 | 34.0 | 24.0 | 2.2 | 7.3 | 29.4 | 21.3 |



and "consumption factor". Table 2 shows the four clinical subtypes of users of bromazepam that were identified from latent class model after users assignment. Proportion of users in each class is given (class distribution) and proportion of each measured variable in each class. Prevalence were similar between both years, therefore we can do the same interpretation. To illustrate, we give here the results for 2009. Class 1 was the most prevalent subtype (61%), characterized by an absence of "doctor shopping" or "pharmacy shopping behaviors" and an absence of overconsumption behavior. The prescriptions of psychotropic drugs came mostly from general practitioners (90%) and were in agreement with the practice guidelines. Class 2 was also a prevalent group (31%), prescriptions stemmed from a general practitioner and were not in agreement with practice guidelines related to benzodiazepines. This class comprised users who could have developed a tolerance after a chronic use. This consumption could be at risk of abuse or dependence in the long term. Class 3 (8%) comprised individuals receiving psychotropic drugs associations with non-agreement of the prescription with the practice guidelines (89%). However this prescription more often stemmed from a specialist (99%). These users may suffer from a resistant or severe mental disorder requiring a larger association of psychotropic drugs not always in agreement with first line guidelines. Lastly, Class 4 was a minority subtype (0.4%). A substantial proportion of these individuals displayed "doctor shopping behavior" (51%), "pharmacy shopping behavior" (96%) and overconsumption behavior (80%). These behaviors are considered as fraudulent and suggest a compulsive use of bromazepam (Wainstein et al., 2011). Therefore, Class 1 was defined as "non-problematic users, Class" 2 as "at risk users," Class 3 as "users with a probable mental disorder" and Class 4 as "compulsive users". Similar proportions were determined in 2008.

### 3.3. Agglomerative Hierarchical Clustering

Using MCA, we retained the first three components that explain 60% of the variance and there was the largest jump between the third and the fourth eigenvalue. Then, AHC was performed on these three components, which were assumed to retain enough information. From dendrogram and statistical index, we retained four groups of users. Table 3 shows the description of the characteristics of the four classes determined by AHC method. For 2009, Class 1 (53%) represented general practitioner prescription, in agreement with practice guidelines. Class 2 (28%) represented general practitioner prescription with at least one association of benzodiazepines: all patients prescriptions were not in agreement with

**Table 3** Description of clusters by agglomerative hierarchical clustering – 2008 and 2009

| | 2008 | | | | 2009 | | | |
|---|---|---|---|---|---|---|---|---|
| | Class 1 | Class 2 | Class 3 | Class 4 | Class 1 | Class 2 | Class 3 | Class 4 |
| *Class repartition (%)* | 52.3 | 28.1 | 17.4 | 2.2 | 53.3 | 27.7 | 16.7 | 2.4 |
| Overconsumption (%) | 0.0 | 0.0 | 0.0 | 51.4 | 0.0 | 0.0 | 0.0 | 50.2 |
| « Doctor shopping » (%) | 0.0 | 0.0 | 0.0 | 17.0 | 0.0 | 0.0 | 0.0 | 18.1 |
| « Pharmacy shopping » (%) | 0.0 | 0.0 | 0.0 | 53.8 | 0.0 | 0.0 | 0.0 | 53.9 |
| General practitioner prescription (%) | 100.0 | 100.0 | 24.2 | 60.0 | 100.0 | 100.0 | 21.2 | 61.7 |
| Not in agreement to the therapeutic class (%) | 0.0 | 100.0 | 55.3 | 66.7 | 0.0 | 100.0 | 52.2 | 69.4 |
| Not in agreement to other classes (%) | 0.0 | 0.0 | 38.8 | 21.8 | 0.0 | 0.0 | 33.6 | 19.5 |



practice guidelines related to the therapeutic class. These two classes differ only on patients prescription and not from patients consumption behavior. Class 3 (17%) was characterized by users with a specialist prescription (79%) and not in agreement with practice guidelines (52% related to benzodiazepines and 34% related to other classes). As with previous classes, there was no overconsumption and doctor or pharmacy shopping behavior. Among consumers of Class 4 (2%), there were 50% of overconsumption, 18% of "doctor shopping" behavior, and 54% of "pharmacy shopping" behavior. Results were similar for both years of the study.

### 3.4. Comparison of Partitions and Stability

Four classes were obtained according to each method, and for both years of the study. These classes allowed to define different levels of consumption. However, there were some noticeable differences. Intra-class inertia was lower for AHC method: 1.2 against 2.7 for LCA. Therefore, AHC partition was more homogeneous. Especially in Classes 1 and 2 (Table 3), there is a very clear and exclusive ranking: prevalence of consumer behaviors are either 0% or 100%. In contrast, partition by LCA is more nuanced: prevalence of consumer behaviors are closest to 10% or 90%.

Tables 4 and 5 represent the confusion matrix between LCA and AHC, showing a different distribution of profiles according to the method. For example in 2009, the four LCA clusters contained 6, 5, 21 and 21, profiles whereas AHC clusters contained 1, 1, 6, and 45 profiles. In AHC, Classes 1 and 2 were more homogeneous, since all individuals have the same profile, but Class 4 was much more heterogeneous with 45 different profiles. For example, in 2009 (Table 5), AHC Class 1 represents 23,838 patients with the same profile: no overconsumption, no doctor shopping, no pharmacy shopping, prescription by a general practitioner, and in agreement with recommendations. AHC Class 2 represents 12,414 patients with no overconsumption, no doctor shopping, no pharmacy shopping, prescription by a general practitioner, in agreement with practice guidelines related to others classes of associated psychotropic drugs but not in agreement with practice guidelines related to the benzodiazepine class. This means that patients of Class 2 may receive simultaneously another benzodiazepine or hypnotic drug.

On the diagonal, there were profiles classified in the same way by both methods and therefore profiles "misclassified" were outside the diagonal. We observe that these misclassified profiles represent few individuals. Indeed, in 2009, half of the

**Table 4** Confusion matrix between LCA partition and AHC partition (data from 2008) with number of responses profiles and number of users

| | | LCA clusters | | | | |
|---|---|---|---|---|---|---|
| | | Class 1 | Class 2 | Class 3 | Class 4 | Total |
| AHC clusters | Class 1 | 1 | 0 | 0 | 0 | 1 |
| | | *21236* | *0* | *0* | *0* | *21236* |
| | Class 2 | 0 | 1 | 0 | 0 | 1 |
| | | *0* | *11435* | *0* | *0* | *11435* |
| | Class 3 | 1 | 2 | 3 | 0 | 6 |
| | | *2218* | *1708* | *3145* | *0* | *7071* |
| | Class 4 | 2 | 4 | 18 | 21 | 45 |
| | | *112* | *291* | *345* | *154* | *902* |
| | Total | 4 | 7 | 21 | 21 | 53 |
| | | *23566* | *13434* | *3490* | *154* | *40644* |

Kappa coefficient = 0.80; Rand indice = 0.89; Jaccard indice = 0.77.

**Table 5** Confusion matrix between LCA partition and AHC partition (data from 2009) with number of responses profiles and number of users

| | | LCA clusters | | | | |
|---|---|---|---|---|---|---|
| | | Class 1 | Class 2 | Class 3 | Class 4 | Total |
| AHC clusters | Class 1 | 1 | 0 | 0 | 0 | 1 |
| | | *23838* | *0* | *0* | *0* | *23838* |
| | Class 2 | 0 | 1 | 0 | 0 | 1 |
| | | *0* | *12414* | *0* | *0* | *12414* |
| | Class 3 | 2 | 1 | 3 | 0 | 6 |
| | | *3246* | *970* | *3234* | *0* | *7450* |
| | Class 4 | 3 | 3 | 18 | 21 | 45 |
| | | *197* | *305* | *369* | *183* | *1054* |
| | Total | 6 | 5 | 21 | 21 | 53 |
| | | *27281* | *13689* | *3603* | *183* | *44756* |

Kappa coefficient = 0.80; Rand indice = 0.88; Jaccard indice = 0.76.

profiles were considered as misclassified but they represented only 12% of the consumers. Most of non-concordant profiles are between Class 3 for LCA and Class 4 for AHC. This represent 18 different profiles, which correspond to 369 users in 2009 (Table 5). It is a mixture of profiles for which it is complicated to tell the difference between them. Indeed, the boundary between a user with a severity of mental disorder (Class 3) and a user with a compulsive behavior (Class 4) is sometimes difficult to define without additional data. Kappa coefficient indicated a good level of concordance between the two clustering methods (80%). Rand and Jaccard index confirms this result, with respectively 88% and 76% of agreement. Results in 2008 were similar.

Stability over time was represented on Table 6 for LCA and on table for AHC. In line with previous results, misclassified profiles were outside the diagonal. In LCA confusion matrix, only 4 over 53 profiles were not in the same class both years, and Kappa coefficient indicated a very good level of concordance (89%). For AHC, all profiles were classified in the same way in 2008 and 2009, so there was a perfect agreement (Kappa = 100%).

**Table 6** Confusion matrix between LCA partition in 2008 and LCA partition in 2009 with number of responses profiles

| | | LCA clusters 2009 | | | | |
|---|---|---|---|---|---|---|
| | | Class 1 | Class 2 | Class 3 | Class 4 | Total |
| LCA clusters 2008 | Class 1 | 4 | 0 | 0 | 0 | 4 |
| | Class 2 | 2 | 4 | 1 | 0 | 7 |
| | Class 3 | 0 | 1 | 20 | 0 | 21 |
| | Class 4 | 0 | 0 | 0 | 21 | 21 |
| | Total | 6 | 5 | 21 | 21 | 53 |

Kappa coefficient = 0.89.

**Table 7** Confusion matrix between AHC partition in 2008 and AHC partition in 2009 with number of responses profiles

| | | AHC clusters 2009 | | | | |
|---|---|---|---|---|---|---|
| | | Class 1 | Class 2 | Class 3 | Class 4 | Total |
| AHC clusters 2008 | Class 1 | 1 | 0 | 0 | 0 | 1 |
| | Class 2 | 0 | 1 | 0 | 0 | 1 |
| | Class 3 | 0 | 0 | 6 | 0 | 6 |
| | Class 4 | 0 | 0 | 0 | 45 | 45 |
| | Total | 1 | 1 | 6 | 45 | 53 |

Kappa coefficient = 1.00.



## 4. DISCUSSION

In this paper, we compared two clustering methods aimed at characterizing drug consumption behavior. Evaluation criteria were the number of clusters obtained by each method, clusters interpretation from a pharmacological perspective, and concordance between the two methods.

The same number of clusters were identified with both methods defining four clinical subtypes of bromazepam users. A good agreement was observed regarding the interpretation of the two first classes. This result indicates an overall consistency between these two clustering methods.

A high stability over time was observed, between 2008 and 2009, for both methods. AHC showed the highest performance regarding this criterion since a perfect concordance was observed: each consumption profile was affected to the same class in 2008 and 2009.

Four different levels of consumptions for each clustering method could be defined. Quantitatively, the consistency between these two methods was high, with 80% of agreement. But a blind interpretation of the results of each method showed several differences. In both cases, a first class was identified corresponding to an adequate consumption of the drug. AHC had the advantage of being more homogeneous, since only one consumer profile was represented: general practitioner prescription, in agreement with practices guidelines, no doctor or pharmacy shopping, and no overconsumption. The same observation can be made for the second class provided by each method. According to AHC partition, only users with a general practitioner prescription and not in agreement with practice guidelines related to benzodiazepines, represented Class 2. With LCA partition, this second class was less homogeneous also including 1% of overconsumption and 7% of prescription not in agreement with other therapeutic classes. Heterogeneity appeared in Class 4 of the AHC with 45 different profiles (23 profiles with LCA). This class of the AHC does not seem to only characterize extreme deviant behaviors but include several types of consumer behavior: 18% of doctor shopping (51% with LCA), 54% of pharmacy shopping (96% with LCA), and 50% of overconsumption (80% with LCA). Indeed, these three variables (doctor and pharmacy shopping, overconsumption) were most relevant to define the patient's behavior while others variables rather characterized the prescription. Moreover, these variables also clearly underline possible deviant behavior. When these three deviant behaviors are jointly observed, such a pattern may be considered an extreme deviant behavior. LCA seems to allow a clearer identification of extreme deviant behavior in comparison to AHC and so, better quantifying the risk of stronger addiction (Wainstein et al., 2011). Non-concordant profiles (grouped in Class 4 with AHC but in class 3 with LCA) illustrate this particular point. Indeed looking more closely at these profiles, it could be noticed that 94% of them displayed only one deviant behavior which does not correspond to an extreme deviant behavior. Moreover, all patients in LCA Class 4 have at least one deviant behavior (only 82% in AHC). Proportion of patients with these three deviant behaviors is higher with LCA than with AHC (27% vs. 5%). This latter information is valuable to answer the need of the health authorities to characterize and distinguish deviant or extreme deviant behavior. Existing tools collect cases of drug abuse or pharmacodependence (notified by health professionals). These clustering methods can provide supplementary information: an estimate of the prevalence of pharmacodependence in real conditions of use.

From a statistical point of view, each of these clustering methods has advantages and limits. A major difference between AHC and LCA is that the latter provides an inferential approach. This means that a statistical model is postulated for the study population (Hagenaars and McCutcheon, 2002) and relies on some hypotheses, particularly local independence. Additionally, there are more formal adjustment criteria to make decisions about the number of clusters, as AIC or BIC criteria. LCA is a probability-based classification, which means cases are classified into clusters based upon membership probabilities (posterior probabilities) estimated directly from the model. There is a notion of error for class assignment with class probability and conditional probability. The greatest benefit of LCA is the opportunity to infer result to another population, from posterior probabilities. However, modal assignment is a limit of this type of model. Indeed, if two posterior probabilities are close for the same patient, this latter may be misclassified. LCA may require large computation times and may not be appropriate if the dataset contains many categorical variables. Also, explanatory variables were binary in our study, but if the number of modalities becomes important, interpretation of conditional probabilities may reveal more difficult.

AHC is a descriptive method that is hypothesis free. In contrast to LCA, decision about the number of clusters is based on graphical choice and pharmacological interpretation and not on a statistical index (as BIC). According to our results, it is a very stable method over time. A posteriori, a sensitivity analysis was performed using K-means algorithm, a nonhierarchical clustering method. For this algorithm, number of classes must be defined a priori. In a first time, the four clusters centers were initialized randomly. Results were unstable: they varied depending on the starting values. In a second step, this algorithm was applied to assess the robustness of the AHC. Cluster centers defined by the AHC were used as initial value. No change was observed compared to AHC partition: classes remain stable by applying a nonhierarchical analysis. This methodology validates stability and homogeneity of AHC. However, this partition defined by AHC concerns a specific sample and there is no notion of statistical inference or probability. As a consequence, these results cannot be generalized and do not allow classification of new individuals. If the aim is to reclassify individuals, an additional step is required after AHC. Indeed, from classes defined by AHC, we can use supervised classification methods on a new sample, as discriminant analysis. This point is an advantage for LCA that is jointly a clustering and classification method.

This study is performed on a single drug, so it is not possible to generalize these results. To continue this work, it would be interesting to apply these methods to other psychotropic drugs, in collaboration with pharmacological experts. This could confirm additional contribution of LCA for defining classes, and especially the identification of extreme behavior. Simulation studies could be a possibility to explore more thoroughly the properties of the two methods.

Several limits inherent to databases could be taken in consideration. A first limitation is representativeness of the sample that excludes farmers and independent professions. These social categories may have a different behavior with regard to the consumption of psychotropic drugs. However, the aim of this study was to compare two methods, and they are applied to the same population. A second limitation is that available variables were related to observed behaviors and not to clinical variables. Definition of explanatory variables can also be a source of error. Data of doctor or pharmacy shopping were censored because we do not have the entire history of each patient. Therefore, there was an underestimation of doctor shopping or pharmacy shopping. Moreover, the CF threshold was taken to be equal to one and this choice is possibly not optimal (Bellanger et al., 2013). Measure of benzodiazepine compliance could be improved to take into account concomitance between dispensing. But it is one of the limits of the database that contains only aggregated data. And variables used for definition of consumer behavior were binary, causing a waste of information. To keep all information, a latent profiles model could be used; it is a latent variable model with categorical latent variable and continuous observed variables. Another axis of research, which would be interesting, would be to apply latent transition analysis, the longitudinal version of LCA. This method could assess evolution of consumer's profiles over time and transitions between latent classes.

To conclude, in our context, we observed that the main interest of AHC is higher stability. It is an attractive method for a descriptive work and provides homogeneous classes. LCA constitutes an inferential approach and provides the possibility of reclassification of individuals using the same model. In a pharmacodependence approach, an interesting matter is to characterize extreme deviant behavior and LCA seems to allow identifying them more accurately.

## REFERENCES

Agrawal, A., Lynskey, M., Madden, P., et al. (2007). A latent class analysis of illicit drug abuse/dependence: Results from the national epidemiological survey on alcohol and related conditions. *Addiction* 102(1):94–104.

Anderberg, M. (1973). *Cluster Analysis for Applications*. New York: Academic Press.

Bellanger, L., Victorri-Vigneau, C., Pivette, J., Jolliet, P., Sébille, V. (2013). Discrimination of psychotropic drugs over-consumers using a threshold exceedance based approach. *Statistical Analysis and Data Mining* 6(2):91–101.

Benzécri, J. (1973). *L'analyse des données*. Paris: Dunod.

Billioti de Gage, S., Begaud, B., Bazin, F., Verdoux, H., Dartigues, J., Peres, K., Kurth, T., Pariente, A. (2012). Benzodiazepine use and risk of dementia: prospective population based study. *BMJ* 27:345–57.

Briot, M. (2006). Office parlementaire d'evaluation des politiques de santé. Rapport sur le bon usage des médicaments psychotropes. Technical report, Assemblée nationale.

Cadet-Taïrou, A., Gandilhon, M., Toufik, A., et al. (2008). Highlights from 2006 and initial observations from 2007. The eighth national report from the recent trends and new drugs system (trend). Technical report, Saint-Denis: OFDT.

Cavalie, P., Richard, N. (2012). Etat des lieux de la consommation des benzodiazepines en France. Technical report, Saint-Denis: ANSM.

Chung, H., Flaherty, B., Schafer, J. (2006). Latent class logistic regression: application to marijuana use and attitudes among high school seniors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169(4):723–743.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational Psychology Measurement* 20:37–46.

Frauger, E., Thirion, X., Chanut, C., et al. (2003). Misuse of trihexyphenidyl (artane®, parkinane®): Recent trends. *Therapie* 58(6):541–547.

Frauger, E., Pradel, V., Natali, F., et al. (2006). Misuse of clonazepam (rivotril®): Recent trends. *Therapie* 61(1):49–55.

Gasquet, I., Négre-Pagés, L., Fourrier, A., et al. (2005). Usage de psychotropes et troubles psychiatriques en france: Résultats de l'étude épidémiologique esemed/mhedea 2000/(esemed) en population générale. *Encephale* 31:195–206.

Hagenaars, J., McCutcheon, A. (2002). *Applied Latent Class Analysis*, 1st ed. Cambridge, UK: Cambridge University Press.

Hubert, L., Arabie, P. (1985). Comparing partitions. *Journal of Classification* 2(1):193–218.

Kendler, K., Ohlsson, H., Sundquist, K., Sundquist, J. (2013). A latent class analysis of drug abuse in a national Swedish sample. *Psychological Medicine* FirstView:1–10.

Latry, P., Martin-Latry, K., Labat, A., Molimard, M., Peter, C. (2010). Use of principal component analysis in the evaluation of adherence to statin treatment: A method to determine a potential target population for public health intervention. *Fundamental & Clinical Pharmacology* 25(4): 528–533.

Lebart, L., Piron, M., Morineau, A. (2000). *Statistique Exploratoire Multidimensionnelle*. 3rd ed. Paris: Dunod.

Magidson, J., Vermunt, J. (2004). *The Sage Handbook of Quantitative Methodology for the Social Sciences*. Thousand Oaks, CA: Sage.

McCutcheon, A. (1987). *Latent Class Analysis*. Newbury Park, CA: Sage.

Nylund, K., Asparouhov, T., Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal* 14(4):535.

Owen, A., Videras, J. (2006). Reconsidering social capital: A latent class approach. *Empirical Economics* 37:555–582.

Reboussin, B., Song, E., Shrestha, A., et al. (2006). A latent class analysis of underage problem drinking: Evidence from a community sample of 16–20 year olds. *Drug and Alcohol Dependence* 83(3):199–209.

Verdoux, H., Lagnaoui, R., Begaud, B. (2005). Is benzodiazepine use a risk factor for cognitive decline and dementia? A literature review of epidemiological studies. *Psychological Medicine* 35:307–315.

Wainstein, L., Victorri-Vigneau, C., Sébille, V., et al. (2011). Pharmacoepidemiological characterization of psychotropic drugs consumption using a latent class analysis. *International Clinical Psychopharmacology* 26(1):54–62.

Ward, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 48:236–244.