Health-policyholder clustering using health consumption

Romain Gauchon, Stéphane Loisel, Jean-Louis Rullière

Noname manuscript No.
(will be inserted by the editor)

Health policyholder clustering using medical consumption

A useful tool for targeting prevention plans

Romain GAUCHON[1,2] · Stéphane LOISEL[1] · Jean-Louis RULLIERE[1]

Received: date / Accepted: date

Abstract On paper, prevention appears to be a good complement to health insurance. However, its implementation is often costly. To maximize the impact and efficiency of prevention plans, plans should target particular groups of policyholders. In this article, we propose a way of clustering policyholders that could be a starting point for the targeting of prevention plans. This two-step method considers mainly policyholder health consumption for classification. The dimension is first reduced using a nonnegative matrix factorization algorithm, producing intermediate health product clusters. Policyholders are then clustered using Kohonen's map algorithm. This leads to a natural visualization of the results, allowing the simple comparison of results from different databases. The method is applied to two real French health insurer datasets. The method is shown to be easily understandable and able to cluster most policyholders efficiently.

Keywords Clustering · Health insurance · Kohonen's map · Nonnegative matrix factorization · Prevention.

1 Introduction

Because it reduces risk, prevention appears to be a useful tool for insurers. However, until recently, European insurance companies have been wary of prevention, which has generally been seen as a marketing product. The first ambitious prevention plan was initiated in 1997 by Discovery in South Africa. It then took until 2016 for a major prevention plan to appear in Europe: the Vitality program (arising from the merger of Discovery and Generali), first launched in Germany, allows policyholders to win points by adopting healthy lifestyles that can be converted into gifts or discount coupons.

There are a number of reasons why private companies might be reluctant to develop prevention. First, it is difficult for insurance companies to achieve high participation rates in prevention plans. Even if high participation rates were achieved,

severe adverse selection could arise: the individuals who are most interested in prevention plans are those with specific health risks [5]. Moreover, policyholders can easily switch between health insurance providers, rendering prevention investment less efficient for insurance companies [23].

It is also difficult to choose who will benefit from the plan. Data protection (regarding health data, in particular) is a touchy subject in Europe. For example, French mutual health insurance companies cannot ask for any medical information when a policyholder takes out a new contract. Therefore, it is impossible to know details about policyholders' health since there is no access to the necessary data. Without access to this information, a supervised method aiming to reconstruct a known partition of the health make-up of policyholders is inconceivable. Consequently, in practice, a useful algorithm to predict health outcomes has to be unsupervised.

In addition, the new European General Data Protection Regulation (GDPR) has limited the possible uses of data. In particular, health data are explicitly considered very sensitive. Article 22 of the GDPR states that private companies cannot take decisions that significantly affect an individual only based on an automated process. As such, the targeting of prevention plans by insurance companies is complicated.

Nevertheless, there are some exceptions to the GDPR, allowing insurance companies to use data to target prevention. First, the GDPR has introduced the notion of individual agreement. If an individual has agreed to a certain use of his personal data for a clearly-defined purpose, the use of these data is allowed. If she does not agree, it is still possible to use these data for regulatory compliance (such as creating financial reserves) or the production of aggregate statistics. The clustering method proposed here purports to meet all of these requirements.

Clustering is frequently used in insurance. For example, credibility theory, which leads to the bonus-malus system, is based on the idea that there exist hidden policyholder clusters (e.g., [10]). Computing classes of policyholders with homogeneous risks can also help define an insurance premium (e.g., Henckaert et al. [22]). Clustering can be used to identify fraud (Derrig and Ostaszewski [16]) and classify risks (Yeo et al. [60]). It has been used to improve risk premium estimation by Verrall and Yaboukov [53]. In the health insurance context, Ghoreyshi and Hosseinkhani classify policyholders using the k-means algorithm [20]; Peng et al. apply clustering algorithms to detect fraud in a health insurance dataset [46]; and Kuo et al. cluster Taiwanese respondents in a health insurance dataset matched to medical information to identify the relationships between illnesses [34].

However, to the best of our knowledge, the clustering of policyholders based on the relationship between claims has not been addressed. Clustering methods in insurance do not thus fully exploit the available data, as they do not use factors such as the individual's treatment program when ill.

A way to capture this kind of relationship consists in pre-processing the data to create a frequency matrix, containing one line for each policyholder and one column for each kind of medical act refunded (such as "pharmacy" or "general practitioner"). However, this matrix might be of high dimension, which notably affects the quality of classic clustering methods due to the curse of dimensionality (e.g., [1]).

The analysis of high-dimensional data has been a prolific research area (e.g., [61], [24], [29]). One classic method of dealing with this issue consists in reducing the dimension before clustering (e.g., [2], or [12]). When the dimension is reduced using a singular value decomposition (SVD), the method is called latent semantic analysis [15]. Of course, other dimension reduction algorithms can be used. For example, Mote et al. [39] use a nonnegative matrix factorization (NMF) algorithm to reduce the dimension and a self-organizing map to cluster brain tumor segmentation. To the best of our knowledge, these kinds of methods have not yet been used for the clustering of health policyholders.

The goal of this paper is to present a new clustering method based only on policyholder health consumption in order to help prevention targeting. The resulting clusters capture particular health profiles. To do this, we use a similar process to that in Mote et al., [39]. The dimension is first reduced with the NMF algorithm, preceding the clustering of policyholders using Kohonen's map. This method is then applied to two real French insurance databases. We demonstrate how to carry out each algorithm step via the detailed analysis of the results obtained for a subpart of one of these databases.

The remainder of the paper is organized as follows: Section 2 describes the different databases used in the present study. Section 3 sets out the NMF and the Kohonen's map algorithms. Section 4 describes the clustering process and the tests that are carried out, as well as the global results, keeping prevention in mind. Examples of a prevention program for psychiatric diseases and falls are provided in Section 5. Lastly, Section 6 discusses the method and proposes some possible extensions.

2 Data presentation

This section aims to present the databases used in this paper and the way they were pre-processed.

Standard health insurance databases Health insurers usually possess two different databases. The first one is called the policyholder database and contains all the information the insurer possesses about the policyholder: age, sex, contract details, contact information, etc. This information is typically used to analyze insurance portfolio profitability.

This database is systematically matched to a second one: the health consumption database. This database contains all the information the health insurer needs to reimburse the policyholder when she buys a health product:[1] the nature (sometimes called the medical act), the date and the amount of the expenditure. The elements in this database, and the nature of the consumption in particular, can vary from one insurer to another: some insurers reimburse a product but others do not, and some insurers have more detailed information than others. The health consumption database can be large: from one million entries for a small mutual

[1] The term health product is used for every item of health expenditure that may be refunded by the insurer (such as GP visits, nights at the hospital, medication and glasses).

health insurance company to over one billion for national health insurance systems. This base may also depend on the national health system. For example, in France, medication for long-term diseases (such as cancer) is fully reimbursed by the national public insurer and does not necessarily appear in these databases.

Figure 1 presents an example of an extract of a standard health consumption database.

Policyholder	Medical act	Date	Public insurer refund	Private insurer refund	Remaining charge	Total payed
1	General practitioner	01/01/2019	16,50 €	8,50 €	- €	25,00 €
1	Pharmacy	01/01/2019	20,00 €	3,00 €	4,00 €	27,00 €
1	Glasses	03/01/2019	3,00 €	200,00 €	187,00 €	390,00 €
1	General practitioner	05/02/2019	16,50 €	8,50 €	- €	25,00 €
1	Osteopathy	06/06/2019	- €	60,00 €	5,00 €	65,00 €
2	General practitioner	04/02/2019	16,50 €	8,50 €	- €	25,00 €
2	Dental cares	08/02/2019	16,10 €	6,90 €	17,00 €	40,00 €

Fig. 1 Example of an extract from a standard heath consumption database

Presentation of the two databases used in this paper The method has been applied to databases from two French private health insurance companies: a collective database (**CB**) and an individual database (**IB**). The CB and IB contain policyholders with top-range and mid-range market contracts, respectively, observed over one year. Moreover, policyholders with zero health consumption are removed.

As the two databases come from different companies, there are some small differences. For example, the CB does not distinguish medical specialists with or without prescriptions. The CB also does not separate fee overruns and the legal copayment from the price of health consumption. On the other hand, the IB does not separate biological tests from blood tests. These differences are taken into account when comparing the two databases in Section 4.4.

pre-processing step For each database, four different splits are carried out (women over 62, women between 16 and 62, men between 16 and 62 and men over 62[2]), producing a total of eight subdatabases. Splitting databases this way helps reduce the heterogeneity between populations and speeds up the process. Figure 2 contains descriptive figures for the eight populations.

We set out the results in detail for the eight subdatabases in Section 4.4. For the sake of conciseness, only the subdatabase of the IB containing women over 62 is used to illustrate the results in Sections 4.1 and 4.2. This subdatabase includes over 500 000 health reimbursements and 160 different health products. The entries that concern extra charges (such as additional fees for night consultations) are dropped, except extra charges for home-care services. We merge identical health products that appear separately due to spelling mistakes. Last, every optical product (such as glasses and contact lenses) is merged into a single optical product, although this does not change the results. After these changes, a typical health consumption database contains around 100 different health products.

[2] The legal retirement age in France is 62.

		Individual database		Collective database	
		Headcount	Average age	Headcount	Average age
Men	-62 years	17153	41	15415	43.7
	+62 years	7949	73,9	12691	71.8
Women	-62 years	38386	42,8	15365	43.4
	+62 years	19727	71,8	13269	74.1

Fig. 2 Databases statistics

To start the process, each subdatabase first has to be transformed into a frequency matrix. The frequency matrix has one column for each different health product (i.e., around 100 columns) and one row for each policyholder. Each column contains the number of consumption of one specific health product. For example, the frequency matrix associated with the extract presented in Figure 1 is given in Figure 3. In this kind of frequency matrix, some columns (such as the column corresponding to the health product "Keratotomy") contain many zeros, while others (such as the one corresponding to "Pharmacy") contain almost no zeros.

Policyholder	General prac	Pharmacy	Glasses	Osteopathy	Dental cares
1	2	1	1	1	0
2	1	0	0	0	1

Fig. 3 Frequency matrix associated with Figure 1

This kind of imbalance can influence the results: the dimension reduction method gives a high weight to frequent health consumption and tends to neglect the less frequent consumption. This can be harmful since less frequent health consumption might be more discriminating and informative than frequent consumption. To tackle this subject, two pre-processing methods have been tested on frequency matrices: the tf-idf method (see [48] for a general presentation) and application of a logarithm function to the frequency matrix[3]. Both methods reduce the weight given to the most frequent medical acts. In the two tested datasets, the clusters of policyholders obtained from the logarithm treatment are more homogeneous. Thus, the results obtained using the tf-idf pre-processing method are not shown in this paper.

3 Algorithms

This section sets out the main algorithms used in this article: the Nonnegative Matrix Factorization and Kohonen's map.

[3] More precisely, if H designs the frequency matrix as described above, the matrix $log(H+1)$ is computed.

3.1 NMF algorithm:

This algorithm is a dimension reduction method. First introduced by Paatero and Tapper [42] and Lee and Seung [37], it is almost unknown in the insurance sector. The only application of which we are aware is that of Nesvijevskaia and Taudau, who announced that they had used this method at the 17th Rencontre MutRé [41].

However, this method is widely used in medicine to analyze the human genome (e.g., [9], [36]) and in text mining to extract features (e.g., [37], [45], [30]). The NMF can also be used as a clustering method (e.g., [33]).

For all $n, m \in \mathbb{N}$, we denote by $\mathcal{NM}(n, m)$ the set of nonnegative matrices with n rows and m columns. Let $n, m, k \in \mathbb{N}$, $V \in \mathcal{NM}(n, m)$. The purpose of the NMF is to find $W \in \mathcal{NM}(n, k)$, $H \in \mathcal{NM}(k, m)$ such that $V \approx WH$.

To do so, a number of algorithms have been proposed (e.g., Lee and Seung ([37], [38]), Brunet et al. [9], Pascual-Montano et al. [43], Badea [4], Kim and Park [31] and Pauca et al. [44]). All except the last algorithm have been tested. The "snmf/l" algorithm, created by Kim and Park, was used to obtain the presented results [4].

This method combines the cost function proposed by Pauca et al. [44] with the concept of sparseness, first introduced in an NMF algorithm by Hoyer [26].

Kim and Park propose to find $\min_{W,H}(\frac{1}{2}[\|V - WH\|_2^2 + \alpha \|H\|_2^2 + \beta \sum_{i=1}^{n} \|W(i,.)\|_1^2])$, with β being a coefficient controlling for the sparseness of W, α a coefficient reflecting H's smoothness, as suggested by Pauca et al. and $\| \ \|_2$ and $\| \ \|_1$ denoting, respectively, the L_2 and the L_1 norms.

It is important to note that the cost function is not convex. Therefore, the snmf/l algorithm is only able to find local optima and is sensitive to the initial values of W and H. The classic way to deal with this is to randomly initialize the two matrices a number of times and then compare the resulting local minima. Following the work of Utsumi [51], this convention has been followed in this paper; however, other methods exist (e.g., [7], [35]).

Starting from a random value, Kim and Park propose the following updating rules to find a local minimum:

$$H_{n+1} = \min_{H \geq 0} \left\| \begin{pmatrix} W_n \\ \sqrt{\beta} e_{1 \times k} \end{pmatrix} H - \begin{pmatrix} V \\ 0_{1 \times n} \end{pmatrix} \right\|_2^2,$$

$$W_{n+1} = \min_{W \geq 0} \left\| \begin{pmatrix} H_{n+1}^T \\ \sqrt{\alpha} I_k \end{pmatrix} W - \begin{pmatrix} V \\ 0_{k \times m} \end{pmatrix} \right\|_2^2,$$

[4] Six different implementations have been tested: those proposed by Lee and Seung (Lee, [37]), Brunet et al. (Brunet, [9]), Pascual-Montano et al. (nsNMF, [43]) and Badea (Offset, [4]) and the two proposed by Kim and Park (snmf/l and snmf/r, [31]). The "snmf/l" algorithm yields among the best results, while being significantly faster. Implementations from the R package "NMF", developed by Gaujoux and Seoighe [19], were used in the analysis presented here.

with n, m being the dimensions of A, k the final dimension, $e_{1\times k}$ a vector of height k containing only 1, and I_k the identity matrix. This method, from Van Banthem and Keenan [52], is called Alternative Nonnegativity constrained Least Squares (ANLS) and guarantees convergence. The stopping criterion is based on the optimality criterion of Karush Kuhn Tucker.

3.2 Kohonen's map algorithm:

Kohonen's map, also called a self-organizing map (SOM), is a clustering method based on neural networks [32].[5]. In this network, every neuron is arranged according to a given topology, usually a two-dimensional grid with square or hexagonal disposition. This way, each neuron has neighboring neurons.

Say that one wishes to classify N policyholders using an n-neuron SOM. To begin, each neuron is assigned a random weight m_i[6]. For each learning iteration t, a random policyholder is chosen. It is then possible to determine the neuron that best represents this policyholder by solving $c = \min_{j\in[\![1,n]\!]} \|x - m_j(t)\|$. The neuron c is called the best machine unit (BMU).

Once the BMU is determined, its weight is adjusted to improve policyholder representativeness: $m_c(t+1) = \alpha(t)(x - m_c(t))$. The coefficient $\alpha(t)$ is the learning rate. This falls over time, so learning is fast at the beginning and meticulous at the end.

To create an influence zone, the weights of the BMU's neighbors are also changed. To do this, it is necessary to define a neighborhood function $h(c, i, t)$. This function falls with the distance between the neuron i and the BMU c. It also falls over time: at the beginning, many neurons are adjusted, while at the end, only few are. Last, the weight of neuron i becomes $m_i(t+1) = m_i(t) = \alpha(t)h(c, i, t)(x - m_c(t))$. This process is repeated several times.

Kohonen's maps have been used a few times in insurance, for example, by Hainaut to classify motorcycle insurance policies [21], or by Brockett et al. to study claims fraud [8].

4 The clustering process

The clustering method presented in this paper contains two steps: the dimension is first reduced, and then, policyholders are clustered. Section 4.1 discusses the dimension reduction step.

[5] The R package "Kohonen", developed by Wehrens et al. ([55]), was used in the analysis presented here.

[6] a well-known alternative is to choose the starting points via a PCA; however, Akinduko et al. show that this is not suitable for non-linear datasets [3]

4.1 Dimension reduction using the NMF algorithm

After pre-processing the health database, a matrix with approximately 100 dimensions is obtained. This is too large for traditional clustering methods, such as k-means, which do not perform very well when there are over 15 dimensions (e.g., [6]).[7] Thus, dimension reduction is needed.

The sNMF/l algorithm is thus applied to the pre-processed frequency matrix to reduce the dimension. After calibrating the algorithm by examining the silhouette, cophenetic correlation and dispersion[8], the final space covers 20 dimensions. This way, two matrices W and H are obtained (see Section 3.1), each with close to 20 000x20 and 20x100 dimensions, respectively. It is important to note that both of them can be interpreted.

The matrix H contains one column for each health product and 20 lines. There are two ways to make this matrix easier to understand. First, it is possible to normalize each row by dividing each coefficient by the sum of all the other row coefficients. This normalized matrix can be used to create a heatmap[9] (see Figure 4). On such a heatmap, a dark case reveals a high normalized coefficient. It shows the way each new dimension is represented by each health product.

This matrix shows that each new dimension can be understood as a health product cluster (HPC). For example, Dimension 7 (row 7 in Figure 4), including "Medical apparatus flat fees", "Orthotics" and "Small medical apparatus", is a medical apparatus HPC. Thus, for the rest of the paper, each new dimension is called an HPC. Each HPC can be easily interpreted.

Furthermore, even though "Technical imagery" looks similar to "Radiology", they do not appear in the same HPC. While a human might have made the mistake of merging these two health products, the algorithm distinguishes them. This shows that the merging applied before the use of this algorithm can substantially affect the final results and thus should be carried out with caution.

In Figures 4 and 5, only the most frequent health products are presented for the sake of clarity. Others, such as "Medical travel", are not shown. However, as shown in sub-Section 4.4, the most important health products are not necessarily the most common: less-frequent health products (such as "Orthoptics") can be significant.

To highlight the role of less-common products and thus specify the meaning of each HPC, it is also possible to normalize H by columns (see Figure 5). If each

[7] For the databases used here, dimension reduction dramatically improves clustering.

[8] These three concepts are quality indicators commonly measured for a clustering. Cophenetic correlation and dispersion aim to measure the stability and were introduced by Brunet et al. [9]. The silhouette was presented by Rousseeuw to test whether individuals are well clustered [49].

[9] The health product "Legal copayment" may be unfamiliar to the reader. In the French health system, many health products are partially reimbursed by the public insurer, "l'Assurance Maladie". The price of health products is fixed by law (for example, a GP consultation costs 25 Euros). However, the public insurer does not refund all of this amount (only 16.5 Euros for GPs) to limit health consumption. Here, we call the 25-16.5 gap the "legal copayment". Moreover, GPs are allowed to charge higher fees that are not covered by the public insurer. The reimbursement of the legal copayment is usually covered by private insurance.

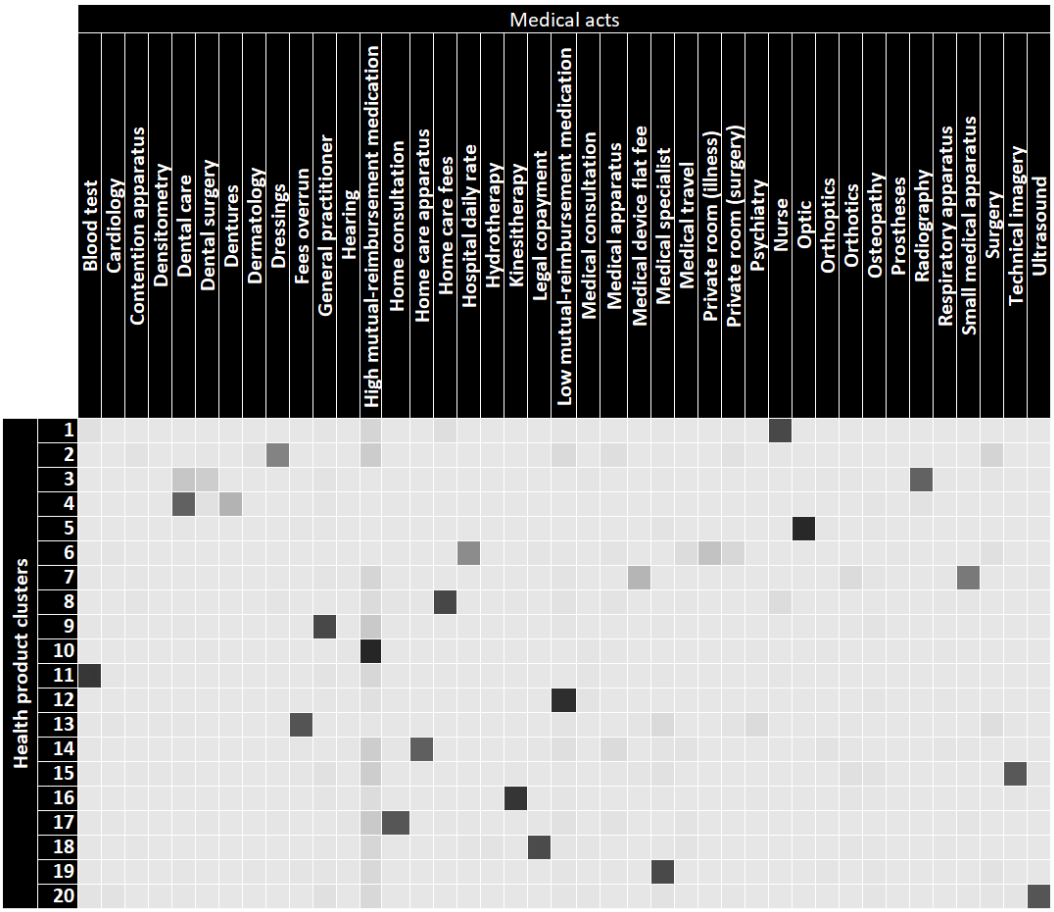

Fig. 4 The horizontal normalization of H. Only the most common health products are shown for clarity.

HPC is seen as a cluster, this normalization shows to which cluster each health product belongs. Thus, a dark case means that a health product closely belongs to the associated HPC.

Using this new heatmap allows a better understanding of each HPC. For example, HPC 3 can now be seen to cover dental surgery, so "Radiography" means "Dental radiography". HPC 15, containing "Densitometry", "Technical imagery", "Orthotics" and "Osteopathy", can be seen as a fracture HPC. As these results come from data on older respondents, this HPC reflects those who have experienced falls.

Once the matrix H is understood, matrix W is easy to read: this contains one line for each policyholder and 20 columns, one for each new dimension. As the latter are interpreted as HPCs, W shows the HPC consumption for each policyholder. The dimension of this matrix is reasonable, allowing the performance of

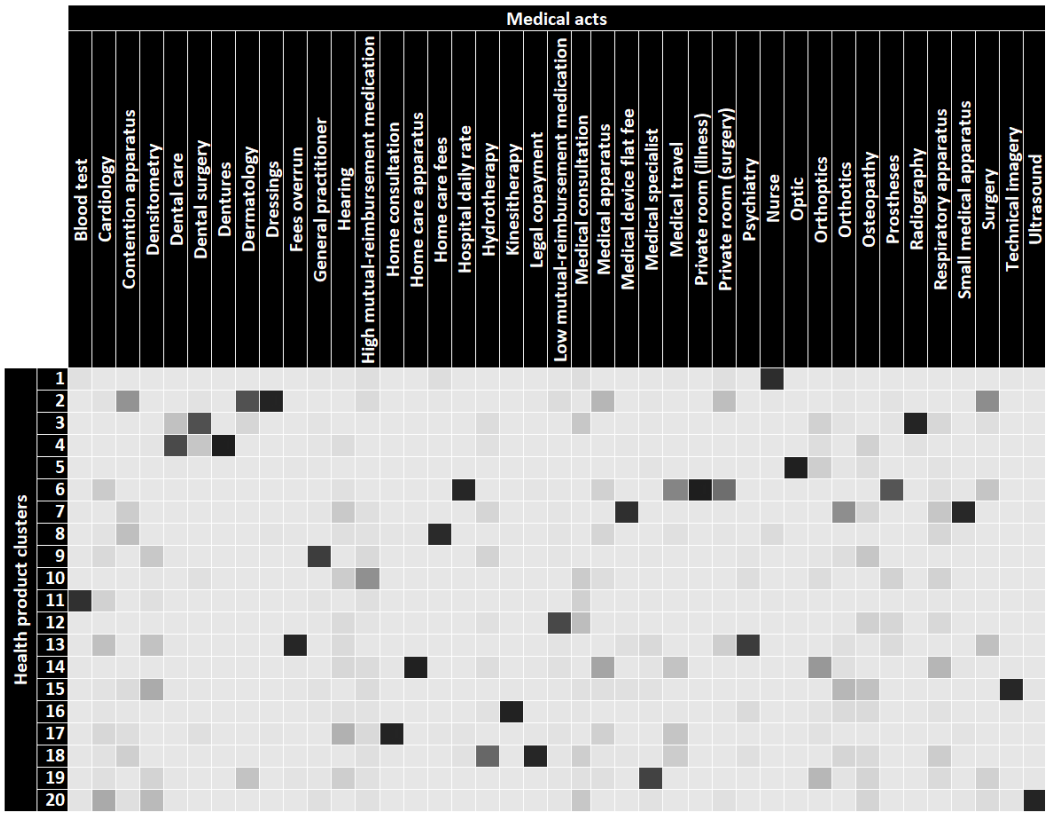

Fig. 5 The vertical normalization of H. This heatmap helps interpret the meaning of each HPC.

clustering. For the following, matrix H is called the HPC matrix, and matrix W is the policyholders matrix.

4.2 Clustering using Kohonen's map

Once the dimension has been reduced, it is possible to cluster policyholders using the policyholder matrix W.

The classic Kohonen's map method with a linear neighborhood function and a learning rate that is decreasing linearly has been used. Gaussian neighborhood functions have also been tested. However, in our datasets, they often produce maps with more empty neurons, although this is not always the case. The map associated with the presented results is given in Figure 6.

In this map, the neurons are disposed on a hexagonal grid with 20 lines (25 neurons per line). The hexagonal grid has been chosen following Kohonen's recommendations [32]. We have chosen a high number of neurons on purpose. Empirically, we might not be able to capture all of the characteristics of the dataset

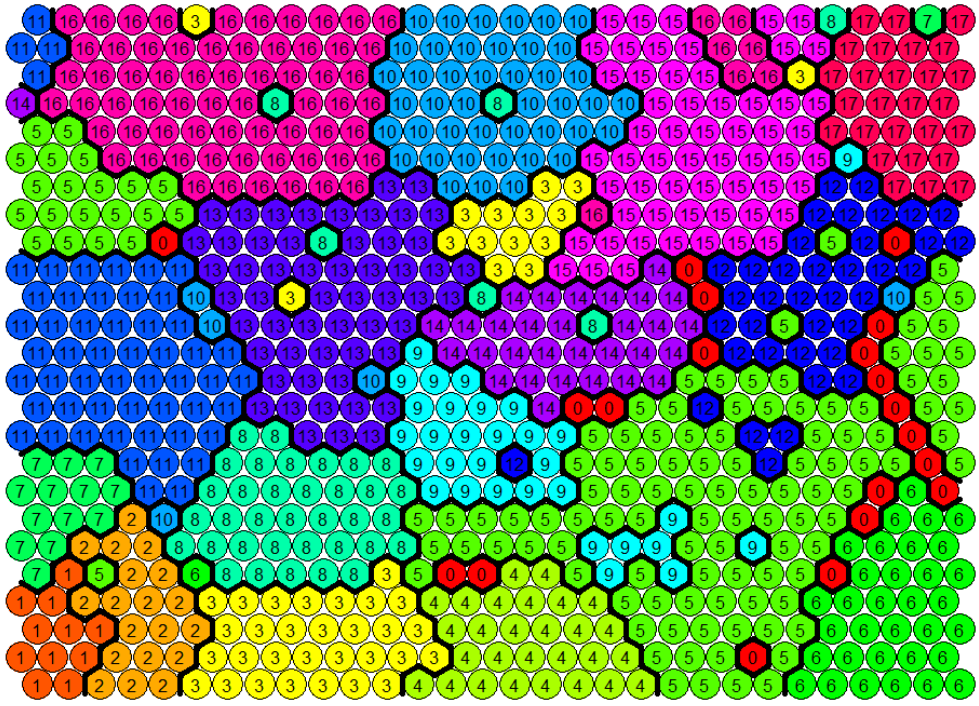

Fig. 6 An example of one of the Kohonen's maps. The associated cluster meanings can be found in Figure 7. Each circle represents a neuron, and each color / number represents a different cluster. The red neurons are empty.

if the number of neurons were too low. Moreover, in this case, the number of clusters is chosen in advance and becomes a parameter. A way to deal with this is to choose a high number of neurons and then to reduce the amount of clusters by grouping similar neurons into the same cluster (for example, by performing a hierarchical agglomerative clustering (HAC)), as first suggested by Murtagh [40]. A high number of neurons allows more flexibility, and the HAC can empirically find a suitable number of clusters and reduce overfitting. After processing the HAC and examining the resulting dendrogram, 17 final clusters are retained.

Before applying the Kohonen's map algorithms, the matrix W is scaled. Usually, the input matrix is scaled by columns. This way, all variables have the same weight. However, empirically and surprisingly, scaling by rows offers much better results on the studied datasets than scaling by columns. Such an operation makes every policyholder comparable: a policyholder with many health consumptions carries the same weight as a policyholder consuming only one health product. This allows the algorithm to focus more on the way a policyholder's health consumption is divided among the different HPCs and less on the overall amount of health consumption.

It is possible to notice in Figure 7 that some of the neurons are empty (the neurons with a 0). One tries to avoid this in traditional approaches, as it makes

no sense to have empty clusters. However, it is acceptable when self-organizing maps are combined with an HAC since neurons are not the final clusters: the HAC merges empty neurons with nonempty neurons. Thus, at the end, there are no more empty clusters, and every final cluster makes sense. Here, empty neurons mainly mark the edge between clusters and low-density areas. To remove empty neurons, we would need to drastically reduce the number of neurons, and the cluster quality would suffer.

The visual size of the clusters in Figure 6 is closely correlated with their real size. Cluster 5 is thus the most populated.

To interpret the clusters, it is possible to calculate the centroid of each class, using the policyholder matrix W. This way, a policyholder of each cluster is computed, representing the whole cluster. This policyholder "type" is described by 20 characteristics, one for each computed HPC. Keeping this in mind, it is easy to interpret each cluster (see Figure 7). Note that a cluster "Radiography" is obtained, which can be read as individuals who are likely to suffer fracture from falls, as the database here covers older respondents.

Policyholders clusters	Blood test	Dental care / denture	Dental surgery	Fees overrun	General Practitioner	Home care apparatus	Home care fees	Home consultation	Hospitalization	Kinesitherapy	Legal copayment	High mutual reimbursement medication	Medical device	Medical specialist (with prescription)	Nurse	Optic	Radiography	Surgery	Ultrasound	Low mutual reimbursement medication	
	Health product clusters																				
1	0,4	0,1	0	0,1	0,3	0,2	1,8	0,6	0,3	0,3	0,3	0,3	0,2	0,2	1,7	0,1	0,1	0,4	0,1	0,4	Home care
2	0,5	0,1	0,1	0,2	0,4	0	1,1	0,1	0,2	0,2	0,2	0,3	0,2	0,3	1,1	0,1	0,2	0,4	0,2	0,4	Nurse home care
3	0,2	0,1	0,1	0,2	0,3	0	0	0	0,1	0,1	0,1	0,4	0	0,4	0,1	0,2	0,1	0,7	0,1	0,4	Surgery
4	0,3	0,1	0,1	0,2	0,3	0,9	0	0,1	0	0,1	0,1	0,4	0,1	0,3	0,1	0,2	0,1	0,1	0,1	0,4	Home care apparatus
5	0,2	0,1	0	0	0,3	0	0	0	0	0	0	0,5	0	0,1	0,1	0	0	0	0	0,2	Everyday care
6	0,2	0,1	0	0,1	0,2	0	0,1	0,8	0,1	0	0,1	0,5	0,1	0,2	0,1	0	0,1	0,1	0,1	0,4	Home consultation
7	0,3	0,1	0,1	0,3	0,2	0,2	1,4	0,4	0,3	1,4	0,2	0,3	0,2	0,3	0,3	0,1	0,2	0,2	0,1	0,4	Home kinesitherapy
8	0,3	0,1	0,1	0,1	0,4	0	0	0	0	0	0,6	0,4	0	0,2	0,1	0	0,1	0	0,1	0,3	Legal copayment
9	0,3	0,1	0	0,3	0,3	0	0	0	0	0	0	0,5	0	0,3	0,1	0	0	0	0,2	0,3	Everyday care with fees overrun
10	0,2	0,1	0	0,1	0,4	0	0	0	0	0	0,1	0,4	0,6	0,2	0,1	0,2	0,1	0,1	0,1	0,3	Medical apparatus
11	0,2	0,2	0,1	0,2	0,4	0	0	0	0	1,3	0,1	0,4	0,1	0,3	0,1	0,1	0,2	0	0,1	0,4	Kinesitherapy
12	0,1	0	0	0,1	0,1	0	0	0	0	0	0	0,1	0	0,1	0	0	0	0	0	0,3	Occasional consumer
13	0,3	0,1	0	0,2	0,4	0	0	0	0	0	0	0,4	0	0,3	0,1	0	0,5	0	0,2	0,3	Radiography
14	0,3	0,1	0	0,1	0,3	0	0	0	0	0	0	0,4	0	0,3	0,1	0	0	0	0,5	0,3	Ultrasound
15	0,2	0,1	0	0,2	0,3	0	0	0	0	0	0,1	0,4	0	0,3	0,1	1,1	0,1	0	0,1	0,3	Optic
16	0,2	0,5	0,6	0,1	0,3	0	0	0	0	0	0	0,4	0,1	0,2	0,1	0,1	0,1	0	0,1	0,3	Dental care
17	0,2	0,1	0	0,1	0,1	0	0	0	0,8	0	0,2	0,3	0,1	0,1	0	0	0	0	0	0,3	Hospitalization

Fig. 7 The center of gravity of the 17 final clusters, calculated using policyholder-matrix coefficients. A high coefficient means that the associated cluster is strongly related to the associated HPC.

A Kohonen's map obtained this way also captures some correlations between clusters. For example, clusters 1, 2 and 7 are close together in Figure 6. It is possible to see in Figure 7 that all of them are "Home-care" clusters.

It is possible to test the consistency of the results by looking at the policyholder profiles in each cluster in terms of variables that were not used in the clustering process, such as age and total medical expenditures (see Figure 8). As expected,

individuals in the "Occasional consumers" cluster do not cost insurers very much, while those in the "Home-care" clusters pose high costs. It is also possible to observe that the average age in comfort-care clusters, such as "Optic" and "Dental care", is higher than the average age in the home-care clusters (e.g., [13]). As such, the method appears to produce consistent results.

Cluster	Meaning	Age	Private Insurer refund	Public insurer refund	Number of policyholders
15	Optic	69	497	284	1419
9	Everyday care with fees overrun	70	145	301	870
12	Occasional consumer	70	62	52	1541
13	Radiography	70	161	407	1393
14	Ultrasound	70	164	350	802
16	Dental care	70	413	448	1771
5	Everyday-day care	71	97	202	4926
8	Legal copayment	71	220	359	970
11	Kinesitherapy	71	307	603	1265
10	Medical apparatus	73	220	399	1069
2	Nurse home care	74	550	794	295
3	Surgery	74	355	532	959
4	Home care apparatus	75	315	533	577
17	Hospitalization	77	1126	277	679
1	Home care	81	792	1389	169
6	Home consultation	81	260	452	826
7	Home kinesitherapy	82	937	1014	196

Fig. 8 This figure represents, for each cluster, the average age and the refund associated with the policyholders belonging to the cluster. Clusters are ordered by average age.

This approach can help choose a particular risk to target using a prevention plan. For example, cluster 17 ("Hospitalization") is expensive for private insurers. It may be of interest to more closely analyze the profile of individuals in this class to target prevention plans. However, medical advice might be needed to fully exploit the results.

This approach is consistent with the European data regulation GDPR. The GDPR distinguishes between two cases. When the policyholder gives her consent to her data being treated for prevention purposes, individuals in each class can be targeted (individuals in cluster 1 can be proposed a specific prevention plan).

Insurance companies do not necessarily have the consent of all policyholders. However, it is still possible to aggregate data to obtain cluster characteristics and thus obtain a general objective characterization (for example, people in home-care clusters are generally over 80 years old, so we can target tertiary prevention plans at those aged over 80 or primary prevention plans at those aged 70 to 80). In our databases, we only know policyholder age, sex and family situation: with more complete information, we could expand the statistical analyses of each cluster.

Please note that the Kohonen's map algorithm may be sensitive to initialization and input data order due to the multi-label context[10] (for example, Appendix 2 shows a map obtained in the same way as Figure 6, changing only the original seed). Empirically, it is possible to see that the cluster meanings are very similar between the different maps. However, some policyholders can change clusters from one map to another. One way to tackle this issue would be to construct a number of different Kohonen's maps based on the same NMF result. It would then be possible to consider policyholders who appear at least once, twice, or every time in a given cluster. We would thus obtain for each policyholder the empirical likelihood of belonging to this cluster, performing a kind of fuzzy-clustering.

4.3 Other tests

Some additional tests have been carried out to better understand the results produced by this method.

Sensibility to infrequent correlations: First, the model's capacity to identify strong but infrequent correlations is tested by adding a randomly-chosen number between 3 and 10 to the "Keratotomy", "Hydrotherapy accommodation" and "Orthoptics" consumption of n random policyholders. "Keratotomy" and "Hydrotherapy accommodation" are consumed only infrequently, while "Orthoptics" is much more common. The NMF algorithm is then rerun. The goal is to identify the minimum n for which the NMF algorithm detects the new correlation. It turns out that only 60 modified policyholders are necessary (out of 20 000 individuals in the database) to detect this correlation.

Challenging this clustering process with a naive approach: Despite all its qualities, the clustering method presented above has to be challenged with a much simpler method. Instead of using this whole process to create a "Hospitalization" cluster (for example), what if one settles on targeting every policyholder with at least one hospitalization health product reimbursement or with hospitalization costs above a certain threshold?

This approach suffers from at least two limitations. First, it is a fully supervised approach: someone has to fix a threshold. This is not a trivial task, and any chosen value could be controversial. Second, in contrast to the process described

[10] An individual can need glasses and have an operation in the same year. Such an individual would then belong to two clusters: the optic cluster and the hospitalization cluster. In this regard, the health sector is linked to a multi-label context.

above, it does not account for correlation, which may lead to inappropriate targeting. For example, it appears that psychiatric diseases are often correlated with hospitalization [18]. However, as discussed in Section 5, it is not optimal to target psychiatric diseases with the same prevention plan as other hospitalizations since they are two very different risks.

In practice, this naive method leads to quite different classes (see Appendix 1 for the detailed results). Most of the time, the NMF-Kohonen method produces clusters with one main type of health consumption, while the clusters obtained with the naive method are more heterogeneous.

Comparison between PCA and NMF: The NMF reduction method is also challenged with the PCA. To do so, the dimension of the IB subdatabases containing only women over 62 has been reduced to 20 dimensions with both methods. Yet, contrary to the pre-processing method presented in Section 2, the optical products (such as "Contact lenses" and "Spectacle lenses") have not been merged together. This way, it is possible to see whether the dimension reduction method is able to make a consistent optic HPC. As shown in Figure 9, the NMF method performs better at this task[11].

Fig. 9 Comparison between ACP and NMF Optic HPC. A darker case means that the health product is detected as closely belonging to the Optic HPC.

In addition, a Kohonen's map has been estimated on the data reduced using the PCA. The R^2 obtained for this clustering is only 0.09. In comparison, the R^2 obtained for the NMF method is usually around 0.25[12].

[11] In fact, almost none of the 20 HPCs made using the PCA offer satisfying consistency.

[12] The R^2 coefficient is given by $R^2 = 1 - \frac{\sum_{i=1}^{k} I_{C_i}}{I_C}$, with I_{C_i} denoting the inertia of the cluster C_i. The inertia has been computed using the cosine similarity, following the recommendation of Huang [28].

The fact that the PCA does not perform well for this task is interesting since this algorithm has proved to be efficient in a much more high dimensional context [27]. However, it is well-known that the classical PCA algorithm is very sensitive to the presence of outliers ([59], [58]), which could explain such poor quality. As shown by Wu et al., the sNMF/l algorithm used in this paper performs better than the PCA in the presence of outliers and noise [57].

Test of the method on a database with known clusters: For a better understanding of the properties of the clustering process, a test on a database with known clusters has been performed. This test is detailed in Appendix 3. It shows that most of the clusters are pure and consistent, even though the largest clusters are usually quite heterogeneous.

4.4 Presentation of the results obtained using the eight subdatabases

In the previous sections, the process was explained and illustrated only using the subdatabase of the Individual Database (IB) containing only women over 62 years old. In this section, the results obtained using the eight subdatabases are presented[13]. Since presenting the results for each subdatabase as precisely as done previously would take too long, a more concise point of view is adopted.

HPC discussion: The HPCs obtained for the eight subdatabases are summarized in Figure 10. In total, 22 HPCs are made in each CB subdatabase, while only 20 are found for each of the IB subdatabases. The HPCs obtained for the IB and the IC are essentially the same, with the main differences being due to the database construction. The method is thus resistant to a change in the database.

However, some differences merit discussion. The HPCs "Respiratory apparatus" and "Home-care apparatus" are more important in the CB than in the IB. This can be due to care refusal: these are expensive products that are better covered by the collective top-range market contract. Moreover, the "Hospitalization" HPC only concerns older people in the individual base, but it concerns everyone in the CB. This is due to the "Legal copayment" HPC, which also contains "Hospitalization" health consumption for younger people in the IB. As the "Legal copayment" health product does not exist in the CB, the HPC becomes "Hospitalization" there.

In both databases, the "Osteopathy", "Orthoptics" and "Psychiatry" HPCs do not concern older people. Osteopathy is a modern practice of which older individuals are less aware, and they prefer going to a physiotherapist instead. It is notable that "Orthoptics"[14] is an HPC on its own. This is a quite narrow health product (in both bases, it is consumed by fewer than 2% of policyholders and represents under 0.12% of total expenditure). Orthoptics mainly covers children,

[13] See Section 2 for a review of the subdatabases.

[14] According to Wikipedia, *"Orthoptics is a profession allied to eye care professions whose primary emphasis is the diagnosis and non-surgical management of strabismus (wandering eyes), amblyopia (lazy eye) and eye movement disorders".*

	Individual database				Collective database			
	Men		Women		Men		Women	
<62years old		Medical specialist (without prescription) Orthoptics Osteopathy Psychiatry			Surgery	Dental care Osteopathy Psychiatry		Orthoptic
>62 years old		Blood test Dental care / Denture / Surgery Drugs Fees overrun General practitioner Kinesitherapy Legal copayment Medical apparatus Medical specialist Nurse Optic Radiography Surgery Ultrasound			Dental care Respiratory apparatus	Blood / Biology test Dental preventive care / Radiology / Denture Drugs Dressings General practitioner Kinesitherapy Hospitalization Medical specialist Nurse Optic Orthotics / medical apparatus Radiography Technical medical procedures Ultrasound		Home care apparatus
	Respiratory apparatus	Home care fees Home consultation Hospitalization		Home care apparatus		Home general practitioner Home care apparatus Personal medical room		Home nurse Orthopedic kinesitherapy

Fig. 10 Summary of the HPCs

which explains why it is an HPC for younger people, with parents paying for their children. Finally, to understand why "Psychiatry" does not appear as an HPC for older people, it is important to underline that in France, dementia is usually treated by neurologists or geriatrician specialists rather than psychiatrists. On the other hand, psychiatric diseases, such as breakdown, often occur before individuals are 30 years old, which explains why this risk appears for younger people. An example of a possible prevention program for the psychiatric cluster is given in Section 5.

Some home-care HPCs are specific to older people, due to dependency. However, it is of interest to note that the "Nurse" HPC typically contains some home-care consumption and exists for both younger and older people.

Last, we can also see that the "Respiratory apparatus" HPC concerns only men, as they are more subject to sleep apnea than women. The "Ultrasound" clusters are found for both men and women, as this is not only used for pregnancy but also for heart, blood and musculoskeletal-system radiography.

Cluster discussion: Once the HPCs have been discussed, it is possible to process the Kohonen's map algorithm. Figure 11 summarizes the interpretation of all clusters obtained for each subdatabase.

First, "Everyday care" and "Occasional consumer" are both clusters with no specific consumption and are combined in the CB dataset. Individuals in these clusters are mostly in good health or refuse treatment.

The HPC cluster meaning analyses are similar to the analysis above (see the comments on "Psychiatry", "Osteopathy", "Orthoptics" and "Respiratory apparatus"). Note that the "Home care" cluster appears in all eight subdatabases.

	Individual database			Collective database		
	Men		Women	Men		Women
<62years old	Blood test	Medical specialist (without prescription) Orthoptics Osteopathy Psychiatry		Blood test Surgery	Dental preventive care / Radiography Hospitalization (with and without personal room) Orthotics / Medical apparatus Osteopathy Psychiatry	Heavy home apparatus Hospitalization Orthoptic
		Dental care Everyday care Home care Kinesitherapy Medical apparatus Legal copayment Optic Radiography Surgery Ultrasound		Dental preventive care / radiography Orthotics / medical apparatus Respiratory apparatus	Biology test Dental care Denture Dressings Everyday care Home care Kinesitherapy Optic Ultrasound	Home care apparatus
>62 years old	Respiratory apparatus	Home care / Kinesitherapy Home consultation Hospitalization Occasional consumer Nurse home care	Everyday care with fees overrun Home care apparatus		Home care device Home general practitioner Hospitalization with personal room Hospitalization without personal room	Orthopedic kinesitherapy Home nurse Hospitalization / Home care

Fig. 11 Interpretation of the final clusters

"Orthotics / Medical apparatus" does not form a cluster for women, and "Home care apparatus" does not form a cluster for men in the CB. For the younger men in the CB, a "Surgery" cluster in addition to the "Dressings" cluster is found, producing two surgery-like clusters. Surprisingly, in both the IB and CB, younger men also have a "Blood test" cluster, which is not the case in the other databases. Last, "Dentures" appears as a cluster only in the CB, mainly due to differences in contract quality.

The analysis of average age and expenditure by cluster also provides considerable information. For example, the average age in the "Dental care" and "Optic" clusters is lower than the average age for older people in both databases. For the younger, the average age in the "Ultrasound" clusters is low, and the cluster contains more female than male policyholders due to motherhood. The "Psychiatry" clusters cover more women than men, which also supports a well-known medical fact (see [56]).

Other findings are more difficult to explain. For young women, the "Kinesitherapy" clusters have a higher average age than those for men, and both are higher than the overall average age. For the IB, the cluster "Medical specialist without prescription" has a lower average age than "Medical specialist with prescription". Last, the average age in the global "Hospitalization" clusters for the older are above the overall average age but below the average age in the "Home care" clusters.

From a more global point of view, the "Everyday care" and "Occasional consumer" clusters are very large. In contrast, there usually exist some very narrow clusters with fewer than 100 policyholders. These usually contain very archetypal consumers and thus can be used to target small prevention plans. To illustrate, in Section 5, we present two prevention plans based on the hospitalization clusters obtained with our approach.

5 Prevention of hospitalization risk.

It is now useful to determine how these results could be used to develop a prevention plan. This section focuses on the hospitalization risk. Working and retired populations are separated since they represent very different risks.

According to medical experts, the hospitalization of policyholders under 62 can be due to very different causes. Figure 11 shows that a "Psychiatry" cluster is created for all 4 subdatabases gathering this population. However, psychiatry (and psychiatric hospitalization) has never been identified as a major risk by French private health insurers. Moreover, the average policyholder in a "Psychiatry" cluster needs much more reimbursement than the average policyholder (e.g., women under 62 in the CB spend 1595 euros on health expenditure, on average, as opposed to 724 euros for the average policyholder). Finally, the "Psychiatry" clusters usually concern a limited number of policyholders (approximately 300 for the same subdatabase as above), limiting the cost of a prevention plan for such a population. Figure 11 also shows that psychiatric clusters are different from hospitalization clusters, underlining the fact that psychiatric illness and hospitalization are two different risks, even though they seem strongly correlated when one looks at the figures [18]. Thus, the clustering process reveals an unexpected risk, which could not be identified by simply applying the naive method presented in Section 4.3 on hospitalization.

However, it is remarkable that once the psychiatric risk has been identified, it is possible to use the simpler method presented in Section 4.3 to target every policyholder with at least one psychiatric consultation. People selected using this simpler method present a similar consumption profile to the one in the "Psychiatry" cluster, even though the number of policyholders targeted is larger. In the case of psychiatric diseases, the method is thus mostly useful to detect the risk or to reduce costs of a prevention plan if one has a tight budget.

Psychiatric risks have never been identified as important for private insurers in France [18]. This is because in France, people suffering from a chronic disease benefit from a special status allowing better refunding from the French public insurer. It is thus commonly believed that chronic diseases mainly impact social security, and private insurers tend to neglect this kind of risk. However, due to comorbidity with other diseases and fee overrun, it appears that this belief is false, at least for psychiatric diseases. The data-driven method presented in this paper has thus revealed something practitioners did not know.

Discussion of these facts with psychiatrists indicates that the people in a "Psychiatry" cluster mainly suffer from severe depression or anxiety disorders and that

they would heavily benefit from a prevention plan. There may be a few people suffering from schizophrenia among them. However, since people with schizophrenia tend to be unemployed, they usually do not benefit from private health insurance.

It also appears that people who see psychiatrists often benefit from sick leave, especially those in the psychiatric cluster, since this cluster concerns people mainly described by a strongly psychiatric-oriented profile. Thus, it is acceptable to suppose that many people in this cluster are actually facing labor disruption. In light of this discussion, an appropriate prevention plan emerges: initiating a return to work. This consists of setting up meetings with the employer and an advisor before the return to work and assisting individuals with setting up a personalized return-to-work schedule to limit anxiety and the risk of relapse. Moreover, it would be helpful to train coworkers in appropriately welcoming back their colleague and reacting in case of relapse. Reducing stigmatization is a key point in the fight against psychiatric diseases.

This example illustrates three points. First, since the whole clustering process is unsupervised, the method proposes a new point of view on data and can reveal unexpected classes of risks. Second, some clusters are small enough to allow a targeted prevention plan. Finally, this clustering process might be a useful tool, but is not enough to target a prevention plan. To design a good prevention program, it is necessary to unite the medical and the actuarial communities.

Concerning retired people, it is possible to see that a cluster "Radiography" is created. According to experts, these clusters gather policyholders who might have suffered a fall. Indeed, when a senior falls, the medical procedure is to systematically look for a fracture in the hospital. This screening phase is very similar to the medical acts consumed by policyholders in this cluster. Thus, some hospitalization can be successfully prevented by prevention against falls, and this cluster seems to be a good target for this kind of prevention program.

Prevention, and in particular, sport has shown to be effective in preventing falls [14]. Examples of statistically efficient programs are Tai Chi Chuan or other physical activity sessions supervised by a physiotherapist.

Some organizations, such as "Siel bleu", are specialized in this field and have proposed programs adapted to various populations. A possible prevention program would thus be to invite policyholders gathered in the "Radiography" cluster to subscribe to a plan managed by this kind of organization. It would mostly consist of practicing adapted physical activity to improve participants' confidence, balance and ability to stand after a fall.

Based on our data-driven study, an insurer could implement several prevention strategies. Specific actions would start with proposing a return to work program with an advisor for workers who are members of the "Psychiatry" cluster and proposing an adapted sports program to members of the "Radiography" cluster.

6 Conclusion

We have presented a method for clustering policyholders based on their health consumption. This first reduces the dimension problem by carrying out a nonneg-

ative matrix factorization. This stage improves the results and helps interpret the clustering by identifying meaningful health product clusters.

In the second stage, policyholders are clustered using Kohonen's maps. This algorithm offers a readable visualization of the results. It allows the simple comparison of clustering carried out using different databases. Moreover, these can be interpreted as different risk clusters. The method has been subjected to a number of tests that reveal its reliability and the quality of the results. Except for the "Everyday care" clusters (composed of occasional consumers or very unique policyholders), most clusters are homogenous and can be used in practice. These clusters are established using common insurance data that are possessed by every health insurer. By constituting clusters, we aggregate the data and so respect the legislation. The method applied here can thus be used to target prevention plans to particular groups of policyholders, and our tests have shown that this process is accurate when we do not have a clear idea of the prevention plans to be instigated. To illustrate this point, an example of a potential prevention program for psychiatric disease, resulting from a talk with a psychiatrist, is proposed. Indeed, thanks to the presented clustering process, psychiatry has been identified as an important risk for private health insurers, which was unknown among practitioners.

There are a number of ways in which the method can be improved. First, this is a mono-label clustering, and multi-label clustering would usually be more appropriate for health risk profiles. This multi-label clustering can be carried out via fuzzy clustering (such as fuzzy c-means) instead of Kohonen's map.

Moreover, the NMF method has considerable advantages but one main disadvantage: due to the initialization requirements, it can be quite slow. To accelerate dimension reduction, other methods may be more appropriate (e.g., word-embedding methods).

The method applied here does not take into account the temporality of health consumption. Knowing whether a policyholder consumes a great deal over a short period or regularly throughout the year could be useful.

The results from this method can be complex and sometimes difficult to interpret. Medical advice for clusters other than psychiatric ones would be useful for understanding the potential scope of this method.

Lastly, this method is not specific to health policyholder clustering and could, for example, also be applied to customer clustering.

Acknowledgements The authors would like to thank Alexandra Barral for useful comments over the duration of this research, and Nabil Rachdi for technical advice. They are also grateful to Addactis in France for providing the data, and too everyone (including two reviewers) who had reread the paper. This research was carried out in the framework of the Chair Prevent'Horizon, supported by the risk foundation Louis Bachelier and in partnership with Claude Bernard Lyon 1 University, Addactis in France, AG2R La Mondiale, G2S, Covea, Groupama Gan Vie, Groupe Pasteur Mutualité, Harmonie Mutuelle, Humanis Prévoyance and La Mutuelle Générale. S. Loisel acknowledges support from the IDR Actuariat Durable sponsored by Milliman Paris, and the DAMI research chair sponsored by BNP Paribas Cardif.

References

1. Aggarwal, C.C., Yu, P.S.: Finding generalized projected clusters in high dimensional spaces, vol. 29. ACM (2000)
2. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: Automatic subspace clustering of high dimensional data for data mining applications, vol. 27. ACM (1998)
3. Akinduko, A.A., Mirkes, E.M., Gorban, A.N.: Som: Stochastic initialization versus principal components. Information Sciences **364**, 213–221 (2016)
4. Badea, L.: Extracting gene expression profiles common to colon and pancreatic adenocarcinoma using simultaneous nonnegative matrix factorization. In: Biocomputing 2008, pp. 267–278. World Scientific (2008)
5. Beaulieu, N., Cutler, D.M., Ho, K., Isham, G., Lindquist, T., Nelson, A., O'Connor, P.: The business case for diabetes disease management for managed care organizations. In: Forum for Health Economics & Policy, vol. 9. De Gruyter (2006)
6. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is "nearest neighbor" meaningful? In: International conference on database theory, pp. 217–235. Springer (1999)
7. Boutsidis, C., Gallopoulos, E.: Svd based initialization: A head start for nonnegative matrix factorization. Pattern Recognition **41**(4), 1350–1362 (2008)
8. Brockett, P.L., Xia, X., Derrig, R.A.: Using kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. Journal of Risk and Insurance pp. 245–274 (1998)
9. Brunet, J.P., Tamayo, P., Golub, T.R., Mesirov, J.P.: Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the national academy of sciences **101**(12), 4164–4169 (2004)
10. Bühlmann, H., Gisler, A.: A course in credibility theory and its applications. Springer Science & Business Media (2006)
11. Cardoso-Cachopo, A.: Improving Methods for Single-label Text Categorization. PdD Thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa (2007)
12. Cheng, C.H., Fu, A.W., Zhang, Y.: Entropy-based subspace clustering for mining numerical data. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 84–93. ACM (1999)
13. Darblade, M.: Analyse de profils de consommation et tarification des futures garanties sur-complémentaire santé. Master's thesis, ISFA (2015)
14. Dargent-Molina, P., Cassou, B.: Prévention des chutes et des fractures chez les femmes âgées. Gérontologie et société **31**(2), 65–78 (2008)
15. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science **41**(6), 391–407 (1990)
16. Derrig, R.A., Ostaszewski, K.M.: Fuzzy techniques of pattern recognition in risk and claim classification. Journal of Risk and Insurance **62**(3), 447–482 (1995)
17. Ding, C., He, X.: K-means clustering via principal component analysis. In: Proceedings of the twenty-first international conference on Machine learning, p. 29. ACM (2004)
18. Gauchon, R., Hermet, J.P.: La psychiatrie: un risque important en assurance santé? (2019)
19. Gaujoux, R., Seoighe, C.: A flexible r package for nonnegative matrix factorization. BMC bioinformatics **11**(1), 367 (2010)
20. Ghoreyshi, S., Hosseinkhani, J.: Developing a clustering model based on k-means algorithm in order to creating different policies for policyholders in insurance industry. International Journal of Advanced Computer Science and Information Technology (IJACSIT) **4**(2), 46–53 (2015)
21. Hainaut, D.: A self-organizing predictive map for non-life insurance. European Actuarial Journal **9**(1), 173–207 (2019)
22. Henckaerts, R., Antonio, K., Clijsters, M., Verbelen, R.: A data driven binning strategy for the construction of insurance tariff classes. Scandinavian Actuarial Journal **2018**(8), 681–705 (2018)
23. Herring, B.: Suboptimal provision of preventive healthcare due to expected enrollee turnover among private insurers. Health Economics **19**(4), 438–448 (2010)
24. Hinneburg, A., Keim, D.A.: Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. pp. 506–517. 25 th International Conference on Very Large Databases (1999)
25. Hinton, G.E., Salakhutdinov, R.R.: Replicated softmax: an undirected topic model. In: Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, A. Culotta (eds.) Advances in Neural Information Processing Systems 22, pp. 1607–1614. Curran Associates, Inc. (2009)

26. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. Journal of machine learning research **5**(Nov), 1457–1469 (2004)
27. Hoyle, D., Rattray, M.: Pca learning for sparse high-dimensional data. EPL (Europhysics Letters) **62**(1), 117 (2003)
28. Huang, A.: Similarity measures for text document clustering. In: Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, pp. 49–56 (2008)
29. Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the thirtieth annual ACM symposium on Theory of computing, pp. 604–613. ACM (1998)
30. Jones, B.W., Chung, W.: Topic modeling of small sequential documents: Proposed experiments for detecting terror attacks. In: Intelligence and Security Informatics (ISI), 2016 IEEE Conference on, pp. 310–312. IEEE (2016)
31. Kim, H., Park, H.: Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics **23**(12), 1495–1502 (2007)
32. Kohonen, T.: The self-organizing map. Proceedings of the IEEE **78**(9), 1464–1480 (1990)
33. Kuang, D., Choo, J., Park, H.: Nonnegative matrix factorization for interactive topic modeling and document clustering. In: Partitional Clustering Algorithms, pp. 215–243. Springer (2015)
34. Kuo, R., Lin, S., Shih, C.: Mining association rules through integration of clustering analysis and ant colony system for health insurance database in taiwan. Expert Systems with Applications **33**(3), 794–808 (2007)
35. Langville, A.N., Meyer, C.D., Albright, R., Cox, J., Duling, D.: Initializations for the nonnegative matrix factorization. In: Proceedings of the twelfth ACM SIGKDD international conference on knowledge discovery and data mining, pp. 23–26. Citeseer (2006)
36. Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al.: Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature **499**(7457), 214 (2013)
37. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788 (1999)
38. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems, pp. 556–562 (2001)
39. Mote, S.R., Baid, U.R., Talbar, S.N.: Non-negative matrix factorization and self-organizing map for brain tumor segmentation. In: Wireless Communications, Signal Processing and Networking (WiSPNET), 2017 International Conference on, pp. 1133–1137. IEEE (2017)
40. Murtagh, F.: Interpreting the kohonen self-organizing feature map using contiguity-constrained clustering. Pattern Recognition Letters **16**(4), 399–408 (1995)
41. Nesvijevskaia, A., Taudou, B.: La data science au service de la prévention santé et prévoyance : nouveaux paradigmes - 17eme rencontre mutré, 14-15 november - nantes. Tech. rep., Malakoff Mederic (2016)
42. Paatero, P., Tapper, U.: Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics **5**(2), 111–126 (1994)
43. Pascual-Montano, A., Carazo, J.M., Kochi, K., Lehmann, D., Pascual-Marqui, R.D.: Nonsmooth nonnegative matrix factorization (nsnmf). IEEE transactions on pattern analysis and machine intelligence **28**(3), 403–415 (2006)
44. Pauca, V.P., Piper, J., Plemmons, R.J.: Nonnegative matrix factorization for spectral data analysis. Linear algebra and its applications **416**(1), 29–47 (2006)
45. Pauca, V.P., Shahnaz, F., Berry, M.W., Plemmons, R.J.: Text mining using non-negative matrix factorizations. In: Proceedings of the 2004 SIAM International Conference on Data Mining, pp. 452–456. SIAM (2004)
46. Peng, Y., Kou, G., Sabatka, A., Chen, Z., Khazanchi, D., Shi, Y.: Application of clustering methods to health insurance fraud detection. In: Service Systems and Service Management, 2006 International Conference on, vol. 1, pp. 116–120. IEEE (2006)
47. Rennie, J.D., Shih, L., Teevan, J., Karger, D.R.: Tackling the poor assumptions of naive bayes text classifiers. In: Proceedings of the 20th international conference on machine learning (ICML-03), pp. 616–623 (2003)
48. Robertson, S.: Understanding inverse document frequency: on theoretical arguments for idf. Journal of documentation **60**(5), 503–520 (2004)
49. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)

50. Settles, B., Craven, M., Ray, S.: Multiple-instance active learning. In: Advances in neural information processing systems, pp. 1289–1296 (2008)
51. Utsumi, A.: Evaluating the performance of nonnegative matrix factorization for constructing semantic spaces: Comparison to latent semantic analysis. In: 2010 IEEE International Conference on Systems, Man and Cybernetics, pp. 2893–2900. IEEE (2010)
52. Van Benthem, M.H., Keenan, M.R.: Fast algorithm for the solution of large-scale non-negativity-constrained least squares problems. Journal of Chemometrics: A Journal of the Chemometrics Society **18**(10), 441–450 (2004)
53. Verrall, R.J., Yakoubov, Y.H.: A fuzzy approach to grouping by policyholder age in general insurance. Journal of Actuarial Practice **7**, 181–204 (1999)
54. Wang, D., Cui, P., Zhu, W.: Structural deep network embedding. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1225–1234. ACM (2016)
55. Wehrens, R., Buydens, L.M., et al.: Self-and super-organizing maps in r: the kohonen package. Journal of Statistical Software **21**(5), 1–19 (2007)
56. W.H.O., et al.: Depression and other common mental disorders: global health estimates (2017)
57. Wu, B., Wang, E., Zhu, Z., Chen, W., Xiao, P.: Manifold nmf with l21 norm for clustering. Neurocomputing **273**, 78–88 (2018)
58. Xu, H., Caramanis, C., Sanghavi, S.: Robust pca via outlier pursuit. In: Advances in Neural Information Processing Systems, pp. 2496–2504 (2010)
59. Xu, L., Yuille, A.L.: Robust principal component analysis by self-organizing rules based on statistical physics approach. IEEE Transactions on Neural Networks **6**(1), 131–143 (1995)
60. Yeo, A.C., Smith, K.A., Willis, R.J., Brooks, M.: Clustering technique for risk classification and prediction of claim costs in the automobile insurance industry. Intelligent Systems in Accounting, Finance and Management **10**(1), 39–50 (2001)
61. Zhang, T., Ramakrishnan, R., Livny, M.: Birch: an efficient data clustering method for very large databases. In: ACM Sigmod Record, vol. 25, pp. 103–114. ACM (1996)

7 Appendix

Appendix 1: Comparing the clusters obtained using the proposed method with those obtained using a very basic approach

Health consumption type	1 - Home care	2 - Nurse home care	3 - Surgery	4 - Home care apparatus	5 - Everyday care	6 - Home consultation	7 - Home kinesitherapy	8 - Legal copayment	9 - Everyday care - fees overrun	10 - Medical apparatus	11 - Kinesitherapy	12 - Occasional consumer	13 - Radiography	14 - Ultrasound	15 - Optic	16 - Dental care	17 - Hospitalization
	Class																
Drugs	409	375	281	300	187	288	319	253	236	252	262	65	253	213	194	199	163
Blood test	74	105	43	47	27	46	52	49	44	42	44	17	52	50	36	38	36
Medical apparatus	155	120	96	184	47	66	207	55	75	154	95	34	64	57	52	84	105
Other	76	31	24	22	7	15	38	56	24	17	26	8	14	14	17	14	23
Nurse / medical auxiliaries	423	60	4	13	2	7	25	2	3	2	3	1	2	2	2	2	1
Surgery	200	235	223	50	11	27	265	22	28	26	34	13	38	28	32	25	167
Dental care	21	88	62	89	67	42	80	86	25	57	107	36	67	81	62	369	41
Home care	319	37	6	11	1	93	112	3	2	3	4	1	2	2	2	2	5
General practitioner	69	124	86	81	46	45	77	84	91	82	101	28	97	83	71	77	33
Hospitalization	347	220	122	102	29	126	495	70	50	51	61	31	68	58	55	44	983
Kinesitherapy	77	48	12	22	3	11	391	8	5	6	257	3	7	5	5	7	8
Optic	113	146	162	181	20	42	109	46	21	170	142	32	32	35	985	138	38
Denture	19	108	74	129	88	60	127	108	29	70	141	28	86	119	70	492	46
Radiography	44	70	44	34	5	24	59	34	9	43	57	6	102	83	33	53	13
Medical specialist	39	58	56	47	11	20	61	30	70	30	46	13	45	38	39	33	18
Total	2386	1824	1296	1311	553	913	2417	905	712	1006	1380	315	931	868	1655	1580	1679
Size	169	295	959	577	4926	826	196	970	870	1069	1265	1541	1393	802	1419	1771	679

Fig. 12 Detailed statistics for all classes.

Health consumption type	Drugs	Blood test	Medical apparatus	Surgery	Dental care	Home care	General practitioner	Hospitalization	Kinesitherapy	Optic	Radiography	Medical specialist	Global
	At least one consumption of :												
Drugs	227	274	302	311	241	332	239	270	290	241	266	265	216
Blood test	40	78	53	60	46	65	45	53	52	45	54	53	39
Medical apparatus	77	92	266	124	88	123	79	118	112	92	98	94	74
Other	18	23	27	29	22	28	20	24	31	23	23	25	18
Nurse / medical auxiliaries	8	10	16	24	6	38	7	11	13	6	8	9	7
Surgery	47	66	81	272	51	111	54	97	85	67	71	85	45
Dental care	92	104	103	106	284	87	103	101	116	113	168	107	92
Home care	11	14	20	23	8	65	9	14	23	8	11	11	11
General practitioner	70	91	90	101	86	84	89	91	105	84	98	97	68
Hospitalization	94	115	167	204	86	196	85	228	144	98	110	110	96
Kinesitherapy	28	34	45	43	33	57	31	37	245	33	40	39	27
Optic	137	153	165	183	169	123	153	192	159	843	169	204	136
Denture	120	139	135	134	367	118	134	131	158	140	219	139	118
Radiography	35	48	50	55	49	45	41	52	63	43	84	53	34
Medical specialist	32	42	45	63	40	40	38	48	52	48	48	72	31
Total	1036	1281	1564	1730	1575	1513	1127	1466	1647	1884	1467	1362	1013
Size	18770	9761	5516	3277	6363	3245	15099	8339	2213	3193	8072	8441	19727

Fig. 13 Consumption if consuming at least some of a particular health product and overall consumption.

Appendix 2: An example of another map obtained from the same data

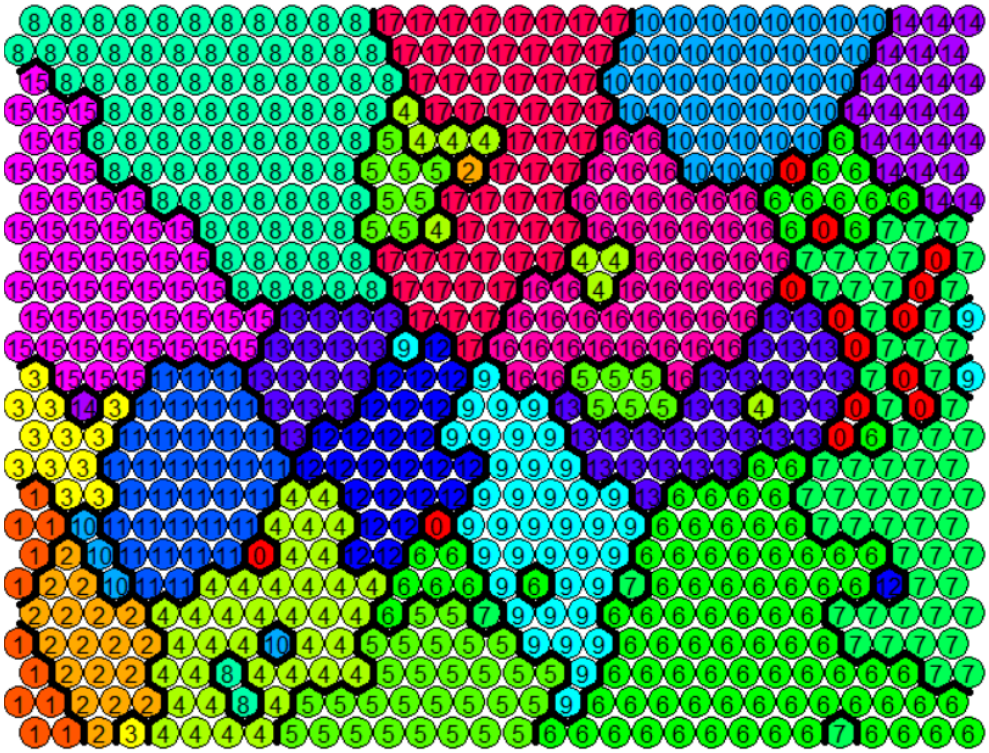

Fig. 14 Self-organizing map obtained from the same data as in Figure 6, using a different seed.

Appendix 3: Test of the algorithm on a database with known clusters

While several tests have been carried out to test the relevance of the results obtained in health insurance databases, it is impossible to compute an objective error metric because the underlying clusters are unknown. To process this kind of test, it is thus necessary to use a dataset coming from another field. For example, the text mining field offers many databases with known clusters. Moreover, it is a common practice in this field to work with a word frequency matrix, making it realistic to apply the NMF/Kohonen process to a text-mining dataset.

The 20-newsgroups dataset is chosen to perform this test. This is a well-known text-mining dataset and has been used for various text-mining tasks, such as word embedding (e.g., [25], [54]), unsupervised clustering (e.g., [17], [28]) and supervised classification (e.g., [50], [47]). The training test dataset has been used, representing 11 293 texts from 20 different newsgroups. For this study, we use the dataset as pre-processed by Cardoso Cachopo [11] (the "no-short" dataset). The objective is to find the original newsgroup of each document.

As text mining is not one of the goals of this paper, the results presented below come from the first run of the algorithm, without attempts to calibrate the model or improve the results. The dimension is first reduced to 60 before clustering, and the frequency matrix is pre-processed using the tf-idf method, which is a common practice in text mining. Since the 20-newsgroups dataset contains 20 different natural clusters, the HAC has been calibrated to obtain 20 different classes.

From the Kohonen map (Figure 15), it is possible to see that clusters 2 and 10 are spread out. Moreover, clusters 6 and 9 seem significantly larger than the others. Their purity score confirms that they are less homogeneous than the other clusters (purity is shown in Figure 17). Except for in these four clusters and cluster 18, purity is acceptable. The overall purity is 62%, and the total entropy[15] is 0.4, which is significantly better than the results obtained by Huang from the same dataset [28], even though we do not aim to achieve a good score.

Comparing Figures 16 and 17, even though the algorithm does not identify all of the documents in a given cluster, the resulting clusters are still reliable. This means that if one wants to identify all of the policyholders with psychiatric medication (for example), this algorithm is not very appropriate. However, if a psychiatric class is identified, it is reliable enough to justify the targeting of a prevention plan.

To summarize, the method produces acceptable results for the 20-newsgroup dataset. Most of the clusters represent a specific newsgroup. However, the method cannot differentiate between very similar newsgroups, such as IBM and Mac computers. This produces large clusters containing most of the documents the method cannot differentiate.

This clustering method is thus able to construct meaningful policyholder clusters. However, large classes (such as the everyday-care cluster) are heterogeneous and should not be used to target prevention plans: they contain policyholders who cannot be differentiated by the algorithm.

[15] The same definition as that of Huang [28] has been used

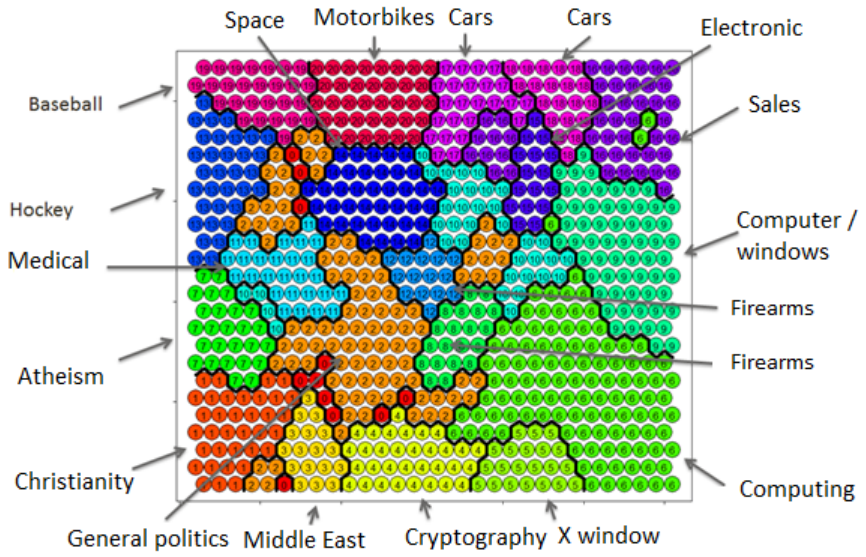

Fig. 15 Kohonen's map using the 20-Newsgroups Dataset. Cluster 10 cannot be constructed.

Newsgroup	Clusters best representing the newsgroup	Document % in the most representative cluster	Document % in the second most representative cluster
Christianity	1	76%	10%
Various politics	2	72%	8%
Middle East politics	2, 3	49%	45%
Cryptography	4	85%	7%
X window system	5, 6	55%	37%
Windows	6, 9	69%	21%
Atheism	7, 1	62%	20%
Firearms	8, 12, 2	38%	32%
IBM computers	10, 6	63%	23%
Digital graphics	6	66%	11%
Medical	12, 2	49%	25%
Various religion	1	44%	30%
Hockey	13	88%	7%
Space	14	77%	8%
Electronic	15, 6	24%	20%
General sales	16	82%	10%
Cars	17, 18	52%	15%
Mac computers	10, 6, 15	50%	22%
Baseball	19	75%	8%
Motorbike	20	87%	4%

Fig. 16 Newsgroup reconstitution capacity

Clusters	Mainly represented newsgroup	Cluster purity	% of total documents clustered in the class
1	Christianity, atheism, various religion	74%	6%
2	Various politics, Middle East politics, Medical, Firearms	27%	11%
3	Middle East politics	96%	3%
4	Cryptography	94%	5%
5	X window system	90%	3%
6	Digital graphics, Windows, X window system, IBM computers, Mac computers, Electronic	24%	15%
7	Atheism	65%	4%
8	Firearms	76%	2%
9	IBM computers, Mac computers, Windows	39%	9%
10	None	20%	3%
11	Medical	91%	3%
12	Firearms	88%	2%
13	Hockey	93%	5%
14	Space	92%	4%
15	Electronic	75%	2%
16	General sales, Mac computers	57%	8%
17	Cars	88%	3%
18	Cars	33%	2%
19	Baseball	96%	4%
20	Motorbikes	93%	5%

Fig. 17 Cluster purity