Use of state sequence analysis for care pathway analysis: The example of multiple sclerosis

Jonathan Roux, Olivier Grimaud, Emmanuelle Leray

Use of state sequence analysis for care pathway analysis: the example of multiple sclerosis

Jonathan Roux[1,2,3], Olivier Grimaud[1,2], and Emmanuelle Leray[1,2,3]

[1] METIS Department, EHESP French School of Public Health, Sorbonne Paris Cité, 15 avenue du Professeur Léon-Bernard - CS 74312, 35043 Rennes, France

[2] UPRES EA 7449 REPERES Pharmacoepidemiology and health services research, Faculty of Medicine/University of Rennes 1 & EHESP French School of Public Health, 2 avenue du Professeur Léon Bernard, 35000 Rennes, France

[3] INSERM CIC-P 1414, CHU of Rennes, 2 Rue Henri le Guilloux, 35000 Rennes, France

Correspondence to:

Jonathan Roux

Département MéTiS - Méthodes quanTitatives en Santé publique

EHESP - École des hautes études en santé publique

15 Avenue du Professeur-Léon-Bernard

CS 74312

35043 Rennes Cedex, France

Email: jonathan.roux@ehesp.fr

Telephone: +33 (0) 2 99 02 29 02

ORCID: 0000-0002-0158-2837

Abstract

The concept of care pathways is increasingly being used to enhance the quality of care and to optimize the use of resources for health care. We here propose an innovative method in epidemiology that is derived from social sciences: state sequence analysis (SSA). This method takes into account the chronology of care consumption and allows for identification of specific patterns. A process for using SSA in the health area is proposed and discussed. The main steps are: data coding, measurement of dissimilarities between sequences (focusing on optimal matching methods and the choice of related costs), and application of a clustering method to obtain a typology of sequence patterns. As an example of its use in the health area, SSA was employed to analyse care pathways of a random sample of patients with multiple sclerosis. This sample has been selected from the main French healthcare database covering the period 2007 to 2013 (n=1 000). A five-cluster typology was obtained which allowed distinction of care consumption groups. Overall, about half of the patients had low care consumption, about one quarter had medium to high consumption, and another quarter had high consumption. We conclude that state sequence analysis is an innovative and flexible methodology that is worth considering in health care research.

Keywords: State sequence analysis; Epidemiology; Care pathways; Multiple sclerosis; Administrative data

1 Introduction

Longer life expectancies and medical advances have led to a steady increase in the number of individuals living with a chronic disease [1]. To take into account the necessary coordination of the multidisciplinary care required for such patients, the concept of a "care pathway" has been defined [2]. This concept is also known as an "integrated care pathway" or a "clinical pathway" and was first used in 1985 [3]. Its aim is to "enhance the quality of care across the continuum by improving risk-adjusted patient outcomes, promoting patient safety, increasing patient satisfaction, and optimizing the use of resources", according to the definition of the European Pathway Association [2,4]. In France, the commonly accepted definition is the organization of overall and continuous care for patients closest of their place of residence, with particular attention being paid to the patient themselves and their own choices [5,6]. Thus, this concept reflects the need for the coordination of health care providers (HCP) (both medical and paramedical), especially between hospital and ambulatory care.

In France, being designated with long disease duration (LDD) status allows individuals with chronic diseases to receive full healthcare coverage. In total, thirty diseases are currently included in the list of LDD, multiple sclerosis (MS) being one of them. The estimated prevalence of MS in France was 151.2/100 000 individuals on December 31, 2012; i.e. a total of about 100 000 individuals nationally [7], with a female to male ratio ranging from 2:1 to 2.5:1 [8]. MS usually starts between 20 and 40 years of age, and most patients typically live with the disease for several decades. The life expectancy of individuals with MS is known to be reduced by 6-7 years compared to the general French population matched for age and gender [9] while the level of disability increases with time. MS symptoms are variable and unpredictable, and each individual's symptoms can change or fluctuate over time. The most common symptoms of this neurological disease are fatigue, difficulty with walking, as well as vision and bladder problems. MS-specific therapeutic options have undergone extensive development since the 2000s, as has the understanding of the importance of comprehensive MS care. The neurologist/general practitioner (GP) dyad is core component of MS care. Indeed, only neurologists are allowed to prescribe MS-specific disease-modifying therapies (DMTs) whereas GPs are most often involved in the day-to-day care. The aim for each patient in this context is "to get the right care provided by the right medical practitioners, at the right time, and at the right place". From a public health point of view, the item "at the right price" can be added. Due to the high cost of DMTs (ranging from €700 to €2 000 per month) and the high need for

care, the financial cost of MS has increased over time on a per patient basis. This increase has resulted in a cost to society of 1.3 billion Euros in France in 2013 [10,11]. In light of the number of individuals in France with MS, the costs of the disease, and the patients' need for complex and multidisciplinary care, the French National Authority for Health is currently working on a recommended care pathway in MS to guarantee appropriate and safe care for all patients. Although an exact overview of care consumption would greatly assist formulating such recommendations, this is currently lacking. Up to now, French studies of care consumption in MS have mainly focused on costs [11,12] and not on the type, amount, and chronology of the care, i.e. the key elements that constitute the care pathway.

The care pathway is unique for each patient due to the choices made by both the HCP and the patient themselves. To date, studies focusing on individual pathways have mainly remained descriptive, without taking into account the possible evolution of care consumption over time [13,14]. At the population level, descriptive data can be summarized and individuals can be categorized according to their level or the amount of care consumption. Such results will be helpful to formulate care recommendations, as well as to anticipate and foresee resources that will be needed in the future. In order to generate such categories, an appropriate methodology taking into account the temporal dimension of care pathways should be used. To our knowledge, the methods used up to now have been derived from informatics, and they have been used to study in-hospital pathways by mining multi-dimensional itemset sequential patterns and symbolic data analysis [15,16]. However, these methods are not commonly used in statistics and epidemiology. Moreover, a requirement for specific knowledge or skills in programming may limit their use in the study of care pathways, including ambulatory care.

In this paper we propose an innovative and easy-to-apply method in epidemiology that is derived from social sciences: the state sequence analysis (SSA). This method is commonly used to identify patterns in work trajectories and to study transitions from education to work [17,18] or career patterns [19,20]. It will allow the chronology of care consumption to be assessed as well as the identification of specific patterns. To our knowledge, this methodology has been used only once in the field of health care, specifically in an attempt to identify disparities in prenatal care pathways [21]. However this study did not provide a clear description of the analytical process. Therefore, the objective of the present manuscript was to highlight the potential and the merits of SSA in the analysis of care pathways and to propose a reference process, from data coding to clustering, through a specific application in regard to multiple sclerosis.

2 Methods

2.1 The study population

The data were obtained from the French permanent sample of health insurance care holders ("Echantillon Généraliste des Bénéficiaires (EGB)" – Agreement n°143 from the French Institute on Health Data). It is a representative random sample of 1/97th of the French national health insurance system ("Système National d'Information Inter-Régimes de l'Assurance Maladie (SNIIRAM)"). EGB gathers data on out-of-hospital reimbursed care consumption, i.e. consultations and home visits with private HCPs, drugs dispensed, LDD status, etc. For each patient, the exact day of execution and type of care are recorded since the date of their inclusion in the database. This sample is updated periodically, i.e. new individuals are included every three months. As the care consumption by these new members is only recorded after their inclusion in the EGB, there is no retrospective data for them. Moreover, a specific characteristic of this database is the fact that patients are not removed from the sample even if they do not receive any care. There are hence no missing values regarding care consumption by patients during follow-up, and only the death of a patient results in its removal from the database. This database was linked to the French hospital discharge database ("Programme de Médicalisation des Systèmes d'Information (PMSI)"), which compiles data on private and public in-hospital care consumption. The database contains the start and end dates of hospital admissions (including day hospital admissions) and their related International Classification of Diseases, 10th version (ICD-10) diagnoses. Outpatient consultations are also included in PMSI. However, the medical speciality is not systematically specified, leading to the exclusion of these consultations from the analysis (as we were only interested in specific specialties, as explained in the next section). Only a small amount of information regarding the patient themselves is available in these two databases, i.e. the year in which they were born; the month and the year of their death; as well as their LDD status and the corresponding year that it started, if applicable.

A four-criteria algorithm has been developed by the French health insurance system for salaried workers (*Caisse Nationale de l'Assurance Maladie des Travailleurs Salariés)* to identify MS cases [7]. Patients are considered to have MS if they meet at least one of the following criteria: having MS LDD status, being prescribed at least one DMT specific for MS (e.g. beta-interferon, glatiramer acetate, fingolimod, or natalizumab), having at least one hospital admission with an ICD-10 diagnosis code G35 [22], or having received at least one disability pension payment for MS. We used these criteria to identify patients with

MS in the EGB database from January 1, 2007 to December 31, 2012. Data regarding these patients were then extracted from 2007 to the end of 2013. In total, one thousand MS patients were identified, of whom 648 were prevalent cases in 2007 (and a further 42 to 94 cases identified annually from 2008 to 2012).

2.2 Data collection

The following elements were included in the care pathways: consultations with GPs (including home visits), consultations with private neurologists and physical medicine and rehabilitation (PMR) physicians (these are considered to be the two main medical specialities involved in MS care in France, in addition to GPs). MS-related hospital admissions (main or related diagnosis of G35, except in-hospital DMT injection) were also considered. Based on the drugs dispensed by pharmacists, the duration of each DMT was determined as the product of the number of packages and the corresponding posology. DMT treatments were included in the analysis if administered for at least three consecutive months, allowing for one month with no medical treatment. As DMTs can only be prescribed by a neurologist, patients receiving DMT without a consultation with a private neurologist were assigned one annual outpatient consultation as long as the treatment was ongoing.

2.3 The SSA method

Using SSA terminology, a care pathway corresponds to a sequence during which successive “states”, i.e. the value taken by the variable at a time t [23], are encountered. The term “sequence” will hence be used to refer to the entire care pathway. In our case study, states are defined as levels of care consumption during each period of time. Thus a sequence may start with a period of high care contact due to disease activity for instance, followed by a period of lower consumption with HCPs related to a period of stable disease. The general principle of SSA is to compare sequences (i.e. individual pathways of patients) with respect to the succession of their component states. In general, SSA can be broken down into three main steps: the coding of data, the measure of similarity (or dissimilarity) between sequences, and the clustering of

sequences. In this work, we focused on dissimilarity-based methods, especially the one that uses optimal matching (OM) algorithms as introduced by Abbott in social sciences [24].

2.3.1 The 1st step of SSA: Identification of states and sequences

The first step of SSA needs to address preliminary questions. First, the study period and the time unit (i.e. day, month, trimester, year, etc.) have to be set. These two items will define the onset and the end of the sequences and the maximum number of states constituting a sequence. Dlouhy et al. recommended using sequences with at least 25 states. They also showed that the number of sequences only has a small effect on the final typology [25]. It is known that missing values can occur at the beginning of some sequences, due to data availability and the data source. Depending on the research question, it can be worth deleting these initial missing states so that the sequences become left-aligned. If the user does not choose to left-align the sequences, some sequences will have a specific state at their beginning coding for the missing value. In this case, the user may opt to split the analysis according to the length of sequences (e.g. short vs. long, or sequences starting in the same year) or simply keep all the sequences in a single analysis. However, the risk of not splitting the sequences into several groups with a homogeneous onset could lead to the creation of a cluster devoted to patients with initial missing states, since their sequences will be more similar. This last scenario may not be meaningful if the main characteristic of this cluster is linked to a delayed start and is not linked to specific patterns of care.
In the present application to MS, inclusion criteria offered the opportunity to identify prevalent cases. Patients were thus not at the same stage of the disease at the start of the study, and there was not a clear indication of the time of the clinical onset of MS. As the disease duration was unknown, the calendar year was considered as the time unit for the analysis. The individual 7-year care pathway was hence obtained by placing the seven annual values of the variable characterizing individual care consumption end-to-end, i.e. one value by year. Sequences could not be left-aligned for individuals entering the study after 2007. To avoid the creation of a cluster devoted to patients with a delayed entry in the cohort and to facilitate understanding of the SSA, we chose to focus the present analysis on patients entering the cohort in 2007 (n=648). The other patients entering the cohort in 2008 or later were analysed separately (n=352) (data not shown).

The next step is to identify a discrete list of values that each state will have to fall into. This is referred as the alphabet in SSA terminology. Choosing the length of the alphabet is important, as it should remain relatively limited. Indeed, the more states there are, the more complicated the final interpretation is likely to be [20]. In our example, a variable characterizing individual annual care consumption was created. It corresponded to the sum of the annual number of consultations with HCPs (i.e. GPs, private neurologists and PMR, and outpatient consultations with neurologists) and MS-related hospital admissions. These annual values of care consumption followed a Poisson distribution with a wide range. We chose to categorize these into five states, using annual quartiles ([0-Q1], [Q1-Q2], [Q2-Q3] and > Q3) plus an extra group with no consumption at all (0). In general (not in our case study due to the characteristics of the database), missing values inside sequences can be considered using an additional state, coded as non-attributed (NA). However, it is recommended to exclude sequences when more than 30% of the states are missing [25].

A sequence ends when the follow-up ends, which induces sequences of different lengths and right-truncated sequences. Depending on the setting, it may be worth considering death as a state of the alphabet. In this case, creating an additional devoted state ("Death") could be an option. However, the use of such a state may favour a high level of similarity between sequences ending with this specific state. In our study, we opted not to create a specific state to code for death, and we decided to stop the sequence of a patient the year after their death. This created right-truncated care pathways and avoided the creation of a specific cluster for patients who had died. Indeed, transitions from any state of care to death were rare in the database. The corresponding substitution costs (described thereafter) would have been higher than the others, and thus explained the gathering of the sequences ending with death in a specific cluster.

2.3.2 The 2nd step of SSA: Measuring dissimilarity between sequences

Once sequences are created, the main component of SSA is the two-by-two comparison of sequences, which requires choosing a dissimilarity measure. Several measures of dissimilarity can be used in SSA. Some were derived from bioinformatics, such as OM, the Levenshtein II distance, and the Hamming

distance. Others were developed in social sciences, such as the longest common subsequence, and the number of matching subsequences [26,27]. In this paper we describe OM, which is the most commonly used method for measuring dissimilarity [28,29], since it can be adjusted for different research issues. The principle of OM is to assign a value to the number of operations necessary to render two sequences strictly similar. As an example, one can consider two sequences made of three consecutive states and that only the last state differs between them. A simple substitution of this state in one sequence will suffice to make the two similar. More operations would be necessary when there is more of a difference between the sequences. OM uses three elementary operations; which are insertion, deletion, and substitution. Insertion corresponds to the addition of a state in a sequence, whereas deletion is the opposite, i.e. the removal of a state. These two operations are joined into a single concept called an "indel", since both operations are warping time. Substitution consists of replacing a state with another one at a time t. To quantify the dissimilarity between two sequences, each of those three operations is assigned a specific cost, which can be constant or which can vary according to the states. The setting of each cost constitutes a crucial step of SSA [30]. Indeed, costs are a major parameter in optimal matching and hence affect clustering results [31]. The dissimilarity between two sequences is then defined as the minimal cost to transform one sequence into the other. This measure therefore depends on the number and the type of elementary operations to transform one sequence into the other and on their costs [29]. To test all of the combinations and to compute the dissimilarity, the Needleman-Wunsch algorithm [32] is used in the main statistical software.

Substitution costs can be chosen either theoretically or empirically. Although in both cases they must respect triangle inequality to ensure that the OM distance and subsequent clustering methods can be computed [29]. The matrix of substitution costs, which is usually symmetrical, contains a value for each substitution of a state by another one and zeroes on the diagonal. Indeed, substituting a state by itself does not change the sequence. If no theoretical costs are available, substitution costs can be estimated empirically using the transition rates observed in the dataset [20,33]. In this case, the cost $c_{i->j/j->i}$ for a substitution of the state s_i by the state s_j, and inversely, equals:

$$c_{i->j/j->i} = 2-p(s_i|s_j)- p(s_j|s_i)$$

where $p(s_i|s_j)$ and $p(s_j|s_i)$ are the probability of observing the transition s_j towards s_i and s_j towards s_i in the dataset, respectively [33]. In regard to the cost of substituting a missing state with another state, it is usually set at 2 when using transition-based costs. The indel cost is set by the user, and most of the time it is fixed to a specific value, although it can also be state-dependant [29], i.e. one value of indel cost for each state of

the alphabet. Attention should be paid to the choice of the indel cost, since indels favour the order of events whereas substitutions preserve contemporaneity [31], as presented in Figure 1. Moreover, indels warp time, while substitutions distort the succession of states in the sequence [31]. Using only indels by setting a very low cost for this operation – this measure is then called the Hamming distance – makes two sequences similar according to their identical states. In contrast, using only substitutions by setting an indel cost greater than the maximum substitution cost – this measure is called the Levenshtein II distance – makes two sequences similar according to their longest common subsequence [31]. (For a full explanation of the process whereby the distance matrix is calculated, see Pollock [34]).
As no theoretical knowledge regarding costs was available for our example, substitution costs were estimated empirically using the observed transition rates. Moreover, we aimed to compare care pathways in terms of the long-term care consumption and we did not want to focus on outstanding events but rather on the whole sequence.

Figure 1 here

2.3.3 The 3rd step of SSA: Clustering sequences

The dissimilarity matrix is used to cluster sequences and to create a typology of patterns according to care consumption. Two main methods can be used for this: agglomerative hierarchical clustering and k-medoids algorithms (similar to k-means). Depending on the predefined costs, the measure used to compare sequences can either be a dissimilarity or a distance (to view axioms of these measures, see Elzinga [35]). It can therefore influence the choice of the criterion to perform agglomerative hierarchical clustering. Ward's criterion is the most used in social sciences [25] with dissimilarity measures [36,37], and it produces the best results in clustering [25]. Unweighted pair-group using arithmetic averages (UPGMA) and weighted pair-group using arithmetic averages (WPGMA) methods can also be considered for agglomerative clustering [25]. In regard to the k-medoids algorithm, the one that we favour is the partitioning around medoids (PAM) algorithm [38]. In order to identify the best clustering, two methods can be applied successively, such as an agglomerative hierarchical algorithm followed by k-medoids [38].
The optimal number of clusters can be chosen either theoretically or empirically. When the Ward's criterion is used, the fall in inertia can permit to set the number of clusters, otherwise quality criteria can

be used. We propose to use a combination of the two main criteria [38]: the weighted average silhouette width (ASWw) [38,39] and the Hubert's C index[40]. The ASWw measures the overall consistency of the clusters. It fluctuates between -1 and 1, a value close to 1 indicating high inter-cluster distances and strong intra-cluster homogeneity. The Hubert's C index, which varies between 0 and 1, reflects the difference between the obtained clustering and the best theoretically possible clustering. Hence a value close to 0 reflects a good clustering. Maximisation of the ASWw and minimisation of Hubert's C index allow for the optimal number of clusters to be set. The choice of these two parameters is in line with a recent social sciences paper, in which three criteria were selected for the partition [41], namely the ASWw, Hubert's C index, and the point biserial correlation (PBC). We chose to forego the PBC as it usually led to the same choice of the number of clusters as the two other criteria.
Here, we used agglomerative hierarchical clustering analysis with Ward's criterion on the dissimilarity matrix to create homogeneous groups. The optimal number of clusters was set using quality parameters. Once the clusters were obtained, each group was described according to individual characteristics and care consumption.

2.3.4 Representing sequences

One major advantage of SSA is the wide option of graphical displaysavailable to describe and communicate results of clustering. The most frequently used graphs are the index plot and the chronogram (also referred to as the state distribution plot) (Figure 2). The index plot is the superposition of each sequence of the set which allows the entire individual longitudinal succession of states of each patient to be seen at once [33]. Hence, the x-axis represents the time (either calendar-based or the duration, depending on the previous choice) and the y-axis represents the id of the sequence, which is usually not represented. By contrast, the individual information level is lost when using the chronogram, synthesizing the information. It displays the general pattern of the whole set of sequences [33], for each time unit, on the x-axis, while the cumulative proportion of patients in the different states is presented on the y-axis. Other types of graphs can be used to represent sequences and results of SSA, such as the mean duration in each state, the sequence composed of the modal state (i.e. the most common state at each time unit), or the most frequent sequences (Supplementary Figure 1).

Figure 2 here

All computational and statistical analyses were performed with R v3.2.3 software [42]. The sequence analysis and the clustering used the *TraMineR* library v1.8-11, the *WeightedCluster* library v1.2, and the *cluster* library v2.0.3 [33,38,43].

3 Results

3.1 Population characteristics

Based on our selection criteria, one thousand MS patients were identified during the whole study period, of whom 648 were identified in 2007.Their characteristics are summarized in Table 1. The majority of these patients were women (71.1%) and the median year of birth was 1963 (ranging from 1914 to1997); that is to say, the patients had a median age of about 44 years at the start of the study.
The 648 patients identified in 2007 were older than the 352 identified in 2008 or later ($p < 0.001$), as they had a median age of 46 years in 2007. Of these patients, 511 (78.9%) had LDD status for MS at the beginning of the follow-up, with a median LDD duration of 6.9 years (ranging from 0.0 to 29.2 years). In total, 55 deaths (8.5%) occurred over the follow-up period, at a median of 63.5 years of age. Over the 7-year study period, 639 patients (98.6%) consulted at least once with a GP and 362 patients (55.9%) consulted at least once with a private neurologist. The number of GP consultations per person-year was 8.0 (7.1for men and 8.3 for women) while the number of consultations with a neurologists was 1.1 (1.0 for men and 1.1 for women).

Table 1 here

3.2 Complete care pathway clustering

To compute the dissimilarity matrix between care pathways, the costs of substitution were based on observed transitions in the dataset of the one thousand patients. Using the complete dataset permitted to

take into account all available information regarding care consumption (Table 2). Indel costs were set at 0.9 to approach the Levenshtein II distance, and thereby compare pathways according to the longest common subsequence. The partitioning of the 648 care pathways led to a typology of five clusters (Figure 3); for the best-quality partitions, the ASWw and HC were equal to 0.238 and 0.102, respectively. This partition can be considered of weak quality.

Among the five clusters, clusters 4 and 5 were predominant, with 180 (27.8%) and 188 (29.0%) patients, respectively (Table 3). Based on the mean duration in each state, these two clusters corresponded to patients with overall high care consumption. Cluster 3 comprised 128 patients (19.8%) with medium consumption. By contrast, patients in cluster 2 (n=124 (19.1%)) had low care consumption, and the 28 patients (4.3%) in cluster 1 had no consumption during nearly half of the follow-up period.

Table 2 here

Figure 3 here

Table 3 here

3.3 Characteristics of the final clustering

The patient characteristics and the care consumption according to the clusters obtained are summarized in Table 4.

As expected, care consumption differed from one cluster to another, and it increased as the cluster number went up. The number of consultations with a GP and a neurologist increased from 2.9 to 14.5 over the study period and from 0.2 to 1.3 per patient-year (Table 4). In cluster 1, the follow-up duration was the shortest ($p < 0.001$), which was probably related to the higher number of deaths. Moreover, the number of patients who used at least one DMT was lower in this cluster than in the other clusters ($p=0.001$). Patients in cluster 2 were significantly younger than those in the other clusters ($p < 0.001$). They were the most frequently treated for MS, with a median treatment duration that lasted about three-quarters of the follow-up period. Furthermore, this cluster comprised a nearly equal proportion of men and women, as opposed to the other clusters which comprised more women. Cluster 3 included a large portion of patients who were prescribed a DMT for most of the duration of the follow-up. Cluster 4 was the oldest one. In cluster 5 – which had the highest consumption – the patients were older, with

significantly more comorbidities than those in other clusters ($p < 0.001$). These patients, who went to the hospital more often than the others, mostly received treatment for MS, although they were treated for a shorter period than those in clusters 2 to 4.

Table 4 here

4 Discussion

The aim of the present study was to provide a reference process and recommendations regarding the use of state sequence analysis to analyse longitudinal care pathways. This process was illustrated by a case-study of the care provided to MS patients in France. Application of the proposed methodology revealed five patterns of consumption for the pathways starting in 2007.A similar clustering was obtained for pathways corresponding to patients with a delayed entry in the cohort (data not shown). All of the clusters were mostly driven by a particular dominating state.

As expected, patients with MS overall appear to be more likely than the general population to consult with a HCP. Indeed, when only considering GPs and neurologists, we observed 8.8 consultations per person-year in our sample. On the opposite, the Organisation for Economic Co-operation and Development reported an average of 6.7 consultations per person-year from 2007 to 2013 in the general population when considering all specialities [44]. Compared to the general population in France, with 4.0 GP consultations per patient-year over the study period, patients with MS had 7.7 consultations per person-year [45]. In accordance with the Canadian studies [46,47], this underscores the importance of GPs in MS care. Overall, about half of the patients had “low” care consumption (clusters 1 to 3), about one quarter had “medium” consumption (cluster 4), and another quarter (cluster 5) had “high” consumption. Based on the detailed results obtained from the clustering, it appears that the cluster with the highest number of patients (i.e. cluster 5, with 29.0 % of the patients) had the highest level of care consumption. This was probably related to the MS itself (i.e. a high number of consultations with a neurologist and a high frequency of DMT use) as well as comorbid conditions (e.g. LDD other than MS). Clusters 2, 3, and 4 appear to reflect care associated with MS, and can be considered to be low to medium levels of care. Cluster 1 was the smallest one with only 4.3% of the patients, and it comprised older patients with a

higher proportion of deaths.

In our opinion, SSA is an innovative method that warrants consideration in the field of care pathways in light of several favourable features. Firstly, this method allows for a holistic approach to pathways, and it therefore considers them as a meaningful conceptual unit [48]. Indeed such an approach permits the coordinated succession of HCPs to be taken into account, which is a characteristic of care pathways. It is opposed to conventional approaches in epidemiology that consider each state independently from one another. The comparison of SSA with other methods, which can be appropriate for the study of a care pathway, such as latent class analysis (LCA), revealed that both methodologies yield similar clustering results [41,49,50]. These studies have highlighted that SSA is easier to use and to compute than LCA and that, unlike LCA, it does not require specification of a model. Another major limitation of LCA is the hypothesis of local independence [41], which cannot be assumed in care pathways since the order in the succession of care consumption is primordial. To overcome this limitation, Mikolai and Lyons-Amos proposed the use of latent class growth models (LCGM) that takes into account the ordering of events, and they reached a conclusion regarding the specifications for the use of each model [50]. However, further work is needed to better evaluate the merits of SSA compared to the latent class model in the field of care pathways.

Secondly, we demonstrated that SSA is a flexible methodology. It can be adapted to different epidemiological contexts in light of the possibility of allowing specification of the measure between sequences amongst others. To transfer this method from social sciences to areas of health care, we used choices based on empirical experiments, namely the effect of costs and death coding. Thus, the choice of splitting and then analysing sequences according to their length avoids the creation of a cluster for patients with delayed entry into the cohort. Similarly, with the choice of terminating the sequence after the death of a patient, we can account for the death of patients without having to create a specific group, which may have occurred with a devoted state.

The data were based on a representative random sample of the French population, which reduces selection bias and which ensures external validity of the results. Moreover, the characteristics (e.g. gender, age, and LDD status) of our study population are close to those of MS patients in France in general [7,11]. The identification of MS patients was based on several criteria, which allows for the detection of low care consumers who may only be identified according to one criterion. In addition, by using administrative databases, we accessed all reimbursed care consumption by patients, independently of self-

reporting. This minimised memory bias and also represented a measure of care consumption as close as possible to the true consumption of the study population.

However, some limitations are due to the database itself. Indeed, EGB does not allow for clear identification of outpatient consultations with hospital-based neurologists, even though we know that MS expert centres are mainly located at French university hospitals. Furthermore, 31% of the hospital consultations in France are with a specialist [51], and only 21.3% of French neurologists were in private practice in 2013 [52]. Therefore, such hospital-based consultations with neurologists were imputed, although the true level of consumption is certainly greater than the assigned level. Indeed, we probably missed the close follow-up at DMT initiation, the monitoring of potential adverse events of DMT, and the follow-up consultations for untreated patients.
Unfortunately, only a few individual characteristics were available to describe the patterns, as the administrative databases do not contain clinical data. A discrepancy analysis (a generalization of the principles of analysis of variance to study the relationship between state sequences and covariates) was performed [37]. The results were not shown because the pseudo-R^2 was too small (< 0.05) regarding the available variables. In addition, the quality of the two typologies we obtained can be considered to be quite weak, which does not allow for assessment of a strong structure in the data. However, this limitation is probably due to the fact that the patients were not at the same stage of the disease during the 2007-2013 period. Moreover care consumption may depend on several factors such as MS relapses or DMT use, for example. To build the states of sequences, we chose to sum the different consultations with health professionals based on equal weights. Even if this operation allowed for an easy to understand variable, specific work has to be undertaken to define weights for each type of consultation. Our work was based on calendar years, and we had no information in regard to the onset of the MS for these patients. This does not allow for left-aligned sequences, and clusters are hence driven by a particular state. These specific characteristics of the EGB database (particularly the lack of clinical data) led to some limitations in our study, although this does not detract from the relevance of SSA in achieving our objective.

This work is only a preliminary draft. It will be extended to the entire French population of patients with MS through use of the French national health insurance information system, i.e. the entire source population from which the EGB is extracted (these analyses are ongoing). It will make for a more accurate examination of the care pathways in a larger population of about 110 000 patients. It will mainly analyse outpatient consultations as well as biological tests, medical imaging examinations, and

paramedical and other specialist care. A new method developed in sequence analysis, called multichannel sequence analysis [34,53], will be used. This method is an extension of the conventional SSA presented in this article, and it allows several thematic sequences for one patient to be considered and for them to be studied simultaneously. Moreover, we are also currently developing algorithms to estimate relevant clinical parameters that affect care consumption, such as motor disability or adherence to treatments, by using parameters described by Hess et al. [54]. Indeed, motor disability is a well-known factor involved in the increase in health care use and costs [55,56] while, by contrast, reliance on DMTs reduces care consumption, such as hospitalizations [57,58]. Comorbidities will also be explored more closely, through the recommendations made by Marrie et al. [59,60]. These parameters may offer opportunities to better characterize the different clusters of the typologies obtained. To overcome the problem of quality of the partition, we will try to identify incident MS cases with an algorithm developed for Canadian administrative databases [61], and thereby study the impact of the onset of MS on care consumption.

In conclusion, SSA is a promising way to study care pathways. For the next stage of this investigation, we intend to compare state sequence analysis with latent class models in the field of epidemiology so as to better determine its potential, strengths, and weaknesses in the area of health care. Further work based on SSA use is needed in the context of care pathways.

Legends

Fig.1 Patterns identified by optimal matching (adapted from Lesnard [31])

Fig.2 Index plot (A) and chronogram (B) of the set of care pathways (n=1 000)

Fig.3 Index plots of clusters obtained for the care pathways of patients identified in 2007 after clustering with indel costs fixed at 0.9 and transition-based substitution costs (n=648)

Supplementary Fig.1 Graphs representing the mean duration in each state (A), the sequence composed of the modal state (i.e. the most common state at each time unit) (B), or the most frequent sequences (C) of the set of care pathways (n=1 000)

Acknowledgments

This work was supported by the French National Agency for Medicines and Health Product Safety (ANSM) through the PEPS platform (Pharmacoepidemiology of Health Products).

Compliance with Ethical Standards

Conflicts of interest: The authors have no conflicts of interest to declare.

References

1. World Health Organization. *Innovative care for chronic conditions : building blocks for actions : global report*. WHO, 2002.

2. Vanhaecht K. *The impact of clinical pathways on the organisation of care processes*. Katholieke Universiteit Leuven, 2007.

3. Zander KS, Bower KA and Etheredge MLS. Nursing case management: blueprints for transformation. Boston, Mass.: New England Medical Center Hospitals, 1987.4. Schrijvers G, Hoorn A van, Huiskes N. The care pathway: concepts and theories: an introduction. *International Journal of Integrated Care* 2012; 12: e192.

5. Féry-Lemonnier E. Les parcours, une nécessité. *Actualité et dossier en santé publique* 2014; 88: 12–15.

6. Haute Autorité de Santé (HAS). *Parcours de soins. Questions/Réponses.* HAS, May 2012.

7. Foulon S, Maura G, Dalichampt M, et al. Prevalence and mortality of patients with multiple sclerosis in France in 2012: a study based on French health insurance data. J Neurol 2017; 264: 1185-1192.8. MS International Federation. Atlas of MShttp://www.msif.org/about-us/advocacy/atlas/ (2015).

9. Leray E, Vukusic S, Debouverie M, et al. Excess Mortality in Patients with Multiple Sclerosis Starts at 20 Years from Clinical Onset: Data from a Large-Scale French Observational Study. *PLoS ONE* 2015; 10: e0132033.

10. Direction de la stratégie, des études et des statistiques (DSES). *Personnes prises en charge pour sclérose en plaques (SEP) en 2015*. Caisse Nationale de l'Assurance Maladie des Travailleurs Salariés (CNAMTS), 2017.

11. Lefeuvre D, Rudant J, Foulon S, et al. Healthcare expenditure of multiple sclerosis patients in 2013: A nationwide study based on French health administrative databases. *Multiple Sclerosis Journal - Experimental, Translational and Clinical* 2017; 3: 2055217317730421.

12. Fromont A, Lehanneur M-N, Rollot F, et al. Cost of multiple sclerosis in France. *Revue Neurologique* 2014; 170: 432–439.

13. Cordesse V, Jametal T, Guy C, et al. Analysis of clinical pathway in changing and disabling neurological diseases. *Revue Neurologique* 2013; 4085: 1–529.

14. Tuppin P, Samson S, Fagot-Campagna A, et al. Care pathways and healthcare use of stroke survivors six months after admission to an acute-care hospital in France in 2012. *Revue Neurologique* 2016; 172: 295–306.

15. Egho E, Jay N, Raïssi C, et al. An Approach for Mining Care Trajectories for Chronic Diseases. In: Peek N, Marín Morales R, Peleg M (eds) *Artificial Intelligence in Medicine*. Springer Berlin Heidelberg, 2013, pp. 258–267.

16. Nuemi G, Afonso F, Roussot A, et al. Classification of hospital pathways in the management of cancer: Application to lung cancer in the region of burgundy. *Cancer Epidemiology* 2013; 37: 688–696.

17. Brzinsky-Fay C. Lost in Transition? Labour Market Entry Sequences of School Leavers in Europe. *European Sociological Review* 2007; 23: 409–422.

18. McVicar D, Anyadike-Danes M. Predicting successful and unsuccessful transitions from school to work by using sequence methods. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2002; 165: 317–334.

19. Anyadike-Danes M, McVicar D. You'll never walk alone: Childhood influences and male career path clusters. *Labour Economics* 2005; 12: 511–530.

20. Biemann T, Datta DK. Analyzing Sequence Data: Optimal Matching in Management Research. *Organizational Research Methods* 2014; 17: 51–76.

21. Le Meur N, Gao F, Bayat S. Mining care trajectories using health administrative information systems: the use of state sequence analysis to assess disparities in prenatal care consumption. *BMC Health Serv Res* 2015; 15: 200.

22. World Health Organization. ICD-10 Version:2016. http://apps.who.int/classifications/icd10/browse/2016/en (2015) Accessed date: 2016-05-31.

23. Studer M. Etude des inégalités de genre en début de carrière académique à l'aide de méthodes innovatrices d'analyse de données séquentielles. Geneva: University of Geneva, 2012.24. Abbott A, Forrest J. Optimal Matching Methods for Historical Sequences. *Journal of Interdisciplinary History* 1986; 16: 471.

25. Dlouhy K, Biemann T. Optimal matching analysis in career research: A review and some best-practice recommendations. *Journal of Vocational Behavior* 2015; 90: 163–173.

26. Elzinga CH. Combinatorial Representations of Token Sequences. *Journal of Classification* 2005; 22: 87–118.

27. Elzinga CH. Sequence Similarity. *Sociological Methods & Research* 2003; 32: 3–29.

28. Abbott A, Tsay A. Sequence Analysis and Optimal Matching Methods in Sociology: Review and Prospect. *Sociological Methods & Research* 2000; 29: 3–33.

29. Studer M, Ritschard G. What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2015; 179: 481–511.

30. Robette N and Bry X. Harpoon or bait? A comparison of various metrics in fishing for sequence patterns. BMS: Bull Sociol Methodol/Bulletin de Méthodologie Sociologique 2012; 116: 5–24.31. Lesnard L. Setting Cost in Optimal Matching to Uncover Contemporaneous Socio-Temporal Patterns. *Sociological Methods & Research* 2010; 38: 389–419.

32. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 1970; 48: 443–453.

33. Gabadinho A, Ritschard G, Müller NS, et al. Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software* 2011; 40: 1–37.

34. Pollock G. Holistic trajectories: a study of combined employment, housing and family careers by using multiple-sequence analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2007; 170: 167–183.

35. Elzinga CH. Distance, similarity and sequence comparison. In: Blanchard P, Bühlmann F, Gauthier J-A (eds) Advances in sequence analysis: theory, method, applications. Vol. 2. Berlin: Springer, 2014, pp.51–73.

36. Batagelj V. Generalized Ward and Related Clustering Problems. In: Bock HH (ed) *Classification and Related Methods of Data Analysis*. Amsterdam, the Netherlands: North-Holland, 1988, pp. 67–74.

37. Studer M, Ritschard G, Gabadinho A, et al. Discrepancy Analysis of State Sequences. *Sociological Methods & Research* 2011; 40: 471–510.

38. Studer M. *WeightedCluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R*. LIVES Working Papers 24, 2013.

39. Kaufman L, Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley, 1990.

40. Hubert L, Levin J. A general statistical framework for assesing categorical clustering in free recall. *Psychological Bulletin* 1976; 83: 1072–1080.

41. Han Y, Liefbroer A and Elzinga C. Comparing methods of classifying life courses: sequence analysis and latent class analysis. Longitudinal and Life Course Studies 2017; 8: 319-341.42. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Core Team. http://www.R-project.org/ (2015) Accessed date: 2015-07-22.

43. Maechler M, Rousseeuw P, Struyf A, et al. *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.3, 2015.

44. OECD Health Statistics. *Health care utilisation*. Organisation for Economic Co-operation and Development (OECD)http://www.oecd-ilibrary.org/social-issues-migration-health/data/oecd-health-statistics/oecd-health-data-health-care-utilisation_data-00542-en (2016) Accessed date: 2016-05-31.

45. Institut de Recherche et Documentation en Economie de la Santé (IRDES). Bases de données Eco-Santéwww.ecosante.fr (2016) 2016-07-29.

46. Marrie RA, Yu N, Wei Y, et al. High rates of physician services utilization at least five years before multiple sclerosis diagnosis. *Multiple Sclerosis Journal* 2013; 19: 1113–1119.

47. Pohar SL, Jones CA, Warren S, et al. Health status and health care utilization of multiple sclerosis in Canada. *The Canadian journal of neurological sciences* 2007; 34: 167–74.

48. Billari FC. Sequence analysis in demographic research. *Canadian Studies in Population* 2001; 28: 439–458.

49. Barban N, Billari FC. Classifying life course trajectories: a comparison of latent class and sequence analysis. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 2012; 61: 765–784.

50. Mikolai J, Lyons-Amos M. Longitudinal methods for life course research: A comparison of sequence analysis, latent class growth models, and multi-state event history models for studying partnership transitions. *Longitudinal and Life Course Studies* 2017; 8: 191–208.

51. Tellier S, De Peretti C, Boisguérin B. *Qui sont les patients des consultations externes hospitalières ?* Direction de la recherche, des études, de l'évaluation et des statistiques (Drees), April 2002.

52. Le Breton-Lerouvillois G. *Atlas de la démographie médicale en France - Situation au 1er Janvier 2013*. Conseil National de l'Ordre des Médecins (CNOM), 2013.

53. Gauthier J-A, Widmer ED, Bucher P, et al. Multichannel sequence analysis applied to social science data. *Sociological Methodology* 2010; 40: 1–38.

54. Hess LM, Raebel MA, Conner DA, et al. Measurement of adherence in pharmacy administrative databases: a proposal for standard definitions and preferred measures. *The Annals of Pharmacotherapy* 2006; 40: 1280–88.

55. Jones E, Pike J, Marshall T, et al. Quantifying the relationship between increased disability and health care resource utilization, quality of life, work productivity, health care costs in patients with multiple sclerosis in the US. *BMC Health Services Research* 2016; 16: 1–9.

56. Kobelt G, Berg J, Lindgren P, et al. Costs and quality of life of patients with multiple sclerosis in Europe. *Journal of Neurology, Neurosurgery & Psychiatry* 2006; 77: 918–926.

57. Halpern R, Agarwal S, Dembek C, et al. Comparison of adherence and persistence among multiple sclerosis patients treated with disease-modifying therapies: a retrospective administrative claims analysis. *Patient preference and adherence* 2011; 5: 73–84.

58. Steinberg SC, Faris RJ, Chang CF, et al. Impact of Adherence to Interferons in the Treatment of Multiple Sclerosis. *Clinical Drug Investigation* 2010; 30: 89–100.

59. Marrie RA, Fisk JD, Stadnyk KJ, et al. Performance of administrative case definitions for comorbidity in multiple sclerosis in Manitoba and Nova Scotia. *Chronic Diseases and Injuries in Canada* 2014; 34: 145–153.

60. Marrie RA, Miller A, Sormani MP, et al. Recommendations for observational studies of comorbidity in multiple sclerosis. *Neurology* 2016; 86: 1446–1453.

61. Marrie RA, Yu N, Blanchard J, et al. The rising prevalence and changing age distribution of multiple sclerosis in Manitoba. *Neurology* 2010; 74: 465–471.

Table 1 Population characteristics according to the year of identification (n=1 000)

	Identified in 2007 n=648	Identified in 2008 or later n=352	Total n=1,000	p-value [a]
No. of women (%)	473 (73.0%)	238 (67.6%)	711 (71.1%)	0.086
Birth year [b]	1961 (1914-1991)	1966 (1924-1997)	1963 (1914-1997)	<0.001
No. of LDD for MS [c]	589 (90.9%)	249 (70.7%)	838 (83.8%)	<0.001
Time since MS-LDD beginning [b,c,d] (years)	12.7 (0.0-34.6)	6.5 (0.2-33.8)	11.6 (0.0-34.6)	<0.001
At least one other LDD [c,e] (%)	186 (28.7%)	92 (26.1%)	278 (27.8%)	0.429
Follow-up [b] (years)	6.9 (0.0-7.0)	3.5 (0.0-6.0)	6.8 (0.0-7.0)	<0.001
No. of deaths (%)	55 (8.5%)	16 (4.5%)	71 (7.1%)	0.029
No. of deaths per person-year (/1,000)	12.9	12.9	12.9	
At least one DMT prescription [f] (%)	368 (56.8%)	171 (48.6%)	539 (53.9%)	0.015
Proportion of follow-up under treatment [b,g] (%)	61.6 (3.3-100.0)	66.0 (7.6-100.0)	63.0 (3.3-100.0)	0.218
Care consumption				
Consultations with a GP [b,h]	43.0 (0.0-350.0)	17.0 (0.0-266.0)	32.0 (0.0-350.0)	<0.001
Consultations with a GP per person-year	8.0	6.5	7.7	
Consultations with a neurologist [b,i]	6.0 (0.0-48.0)	2.0 (0.0-19.0)	4.0 (0.0-48.0)	<0.001
Consultations with a neurologist per person-year	1.1	1.0	1.1	
At least one MS-related hospital admission [j] (%)	307 (47.3%)	166 (47.2%)	473 (47.3%)	1.000
MS-related hospital admissions per person-year	0.24	0.31	0.26	

[a] $P < 0.05$ comparing patients identified in 2007 and patients identified in 2008 or later using the Kruskal-Wallis, Pearson's chi-square, or Fisher's exact test if needed.
[b] Median (minimum-maximum).
[c] LDD: long disease duration.
[d] Calculated at the date of the most recent information (i.e. Dec. 31, 2013 or the date of the patient's death).
[e] LDD other than MS.
[f] DMT: disease-modifying therapy.
[g] For patients with at least one DMT use.
[h] GPs: general practitioners, Number of consultations and home visits, total per patient from 2007 to 2013.
[i] Number of consultations, home visits, and imputed outpatient consultations, total per patient from 2007 to 2013.
[j] MS-related hospital admissions, total per patient from 2007 to 2013 (except monthly DMT injections).

Table 2 Matrix of observed transition rates in the dataset (n=1 000)

	->0	**->]0;Q_1]**	**->]Q_1;Q_2]**	**->]Q_2;Q_3]**	**-> >Q_3**
0 ->	0.490	0.232	0.105	0.108	0.064
]0;Q_1] ->	0.069	0.453	0.313	0.111	0.054
]Q_1;Q_2] ->	0.015	0.245	0.402	0.258	0.080
]Q_2;Q_3] ->	0.009	0.104	0.304	0.345	0.238
>Q_3 ->	0.007	0.020	0.091	0.223	0.658

Table 3 The mean duration in years (part of total duration) in each state among the clustering of patients identified in 2007 (n=648)

Cluster state	n (%)	0	[0–Q1]	[Q1–Q2]	[Q2–Q3]	> Q3
1.1	28 (4.3%)	*3.46 (49.4%)*	1.18 (16.9%)	0.14 (2.0%)	0.64 (9.1%)	0.93 (13.3%)
1.2	124 (19.1%)	0.31 (4.4%)	*4.06 (58.0%)*	1.86 (26.6%)	0.55 (7.9%)	0.19 (2.7%)
1.3	128 (19.8%)	0.09 (1.3%)	1.39 (19.9%)	*4.07 (58.1%)*	1.20 (17.1%)	0.16 (2.3%)
1.4	180 (27.8%)	0.13 (1.9%)	0.57 (8.1%)	1.76 (25.1%)	*2.95 (42.1%)*	0.95 (13.6%)
1.5	188 (29.0%)	0.10 (1.4%)	0.24 (3.4%)	0.45 (6.4%)	1.37 (19.6%)	*4.71 (67.3%)*

The predominant state for each cluster is indicated in italics.
The yearly consumption was classified into 5 groups by quartile of distribution: no consumption (0), less than the first quartile ($[0-Q1]$), between the first quartile and the median included ($[Q1-Q2]$), between the median and the third quartile included ($[Q2-Q3]$), and greater than the third quartile ($>Q3$).

Table 4 Characteristics of patients identified in 2007 by clusters (n=648)

Characteristics	Cluster 1.1 n=28	1.2 n=124	1.3 n=128	1.4 n=180	1.5 n=188	p-value [a]
No. of women (%)	23 (82.1%)	72 (58.1%)	92 (71.9%)	142 (78.9%)	144 (76.6%)	<0.001
Birth year [b]	1958.5 (1930-1989)	1965.5 (1929-1991)	1960 (1923-1988)	1957.5 (1914-1986)	1959 (1916-1987)	<0.001
No. of LDD for MS [c]	24 (85.7%)	110 (88.7%)	120 (93.8%)	167 (92.8%)	168 (89.4%)	0.373
Time since MS-LDD beginning [b,c,d] (years)	12.4 (1.0-26.2)	12.6 (1.0-28.6)	13.0 (3.0-34.6)	12.1 (0.6-34.0)	12.8 (0.0-34.3)	0.629
At least one other LDD [c,e] (%)	9 (32.1%)	19 (15.3%)	32 (25.0%)	48 (26.7%)	78 (41.5%)	<0.001
Follow-up [b] (years)	6.5 (0.0-7.0)	6.9 (0.0-7.0)	6.9 (0.0-7.0)	7.0 (0.0-7.0)	7.0 (0.0-7.0)	<0.001
No. of deaths (%)	6 (21.4%)	4 (3.2%)	4 (3.1%)	26 (14.4%)	15 (8.0%)	<0.001
No. of deaths per person-year (/1,000)	37.4	4.7	4.6	23.4	11.8	
At least one DMT prescription [f] (%)	6 (21.4%)	78 (62.9%)	74 (57.8%)	95 (52.8%)	115 (61.2%)	0.001
Proportion of the follow-up under treatment [b,g] (%)	14.3 (3.6-59.0)	75.5 (3.3-100.0)	76.5 (4.4-100.0)	60.2 (3.3-100.0)	51.8 (4.4-100.0)	<0.001
Care consumption						
Consultations with a GP [b,h]	19.0 (0.0-73.0)	18.5 (0.0-51.0)	32.0 (10.0-54.0)	48.0 (0.0-96.0)	89.0 (31.0-350.0)	<0.001
Consultations with a GP per person-year	4.1	2.9	4.7	7.5	14.5	
Consultations with a neurologist [b,i]	0.0 (0.0-8.0)	6.0 (0.0-23.0)	7.0 (0.0-35.0)	5.0 (0.0-38.0)	7.0 (0.0-48.0)	<0.001
Consultations with a neurologist per person-year	0.2	0.9	1.0	1.1	1.3	
At least one MS-related hospital admission [j] (%)	8 (28.6%)	61 (49.2%)	44 (34.3%)	84 (46.7%)	110 (58.5%)	<0.001
MS-related hospital admission per person-year	0.14	0.23	0.09	0.25	0.35	

[a] $P < 0.05$ comparing the 5 groups by the Kruskal-Wallis, Pearson's chi-square test, or Fisher's exact test.
[b] Median (minimum-maximum).

[c] LDD: long disease duration.
[d] Calculated at the date of the most recent information (i.e. Dec. 31, 2013 or date of the patient's death).
[e] LDD other than MS.
[f] DMT: disease-modifying therapy.
[g] For patients with at least one DMT use.
[h] GPs: general practitioners, Number of consultations and home visits, total per patient from 2007 to 2013.
[i] Number of consultations, home visits, and imputed outpatient consultations, total per patient from 2007 to 2013.
[j] MS-related hospital admissions, total per patient from 2007 to 2013 (except monthly DMT injections).

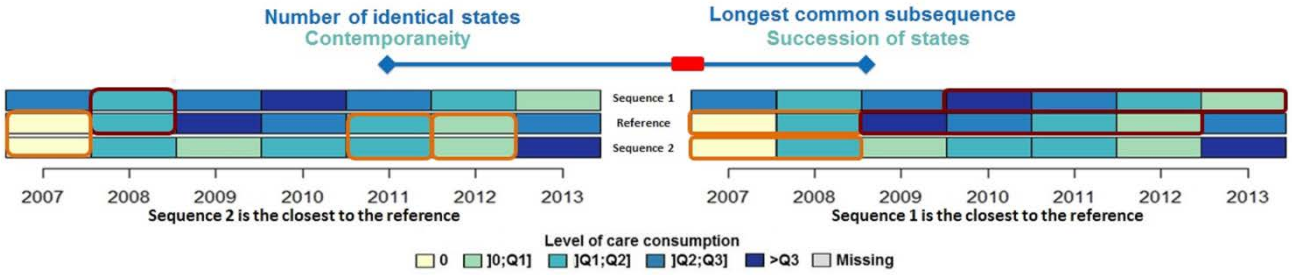


Fig.1 Patterns identified by optimal matching (adapted from Lesnard [31])

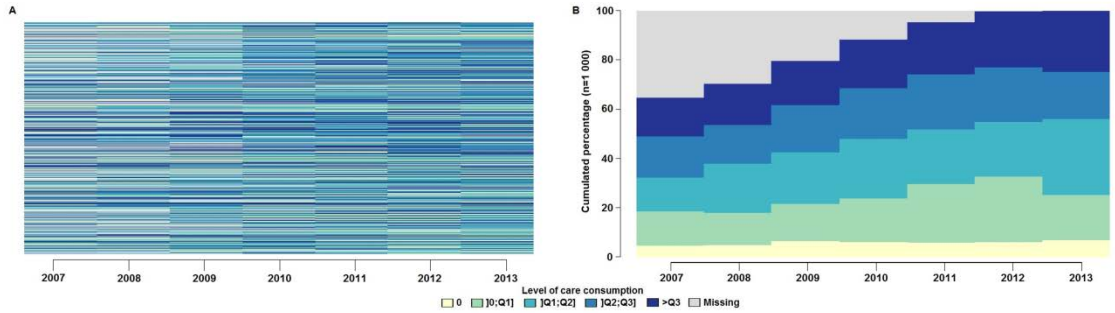


Fig.2 Index plot (A) and chronogram (B) of the set of care pathways (n=1 000)

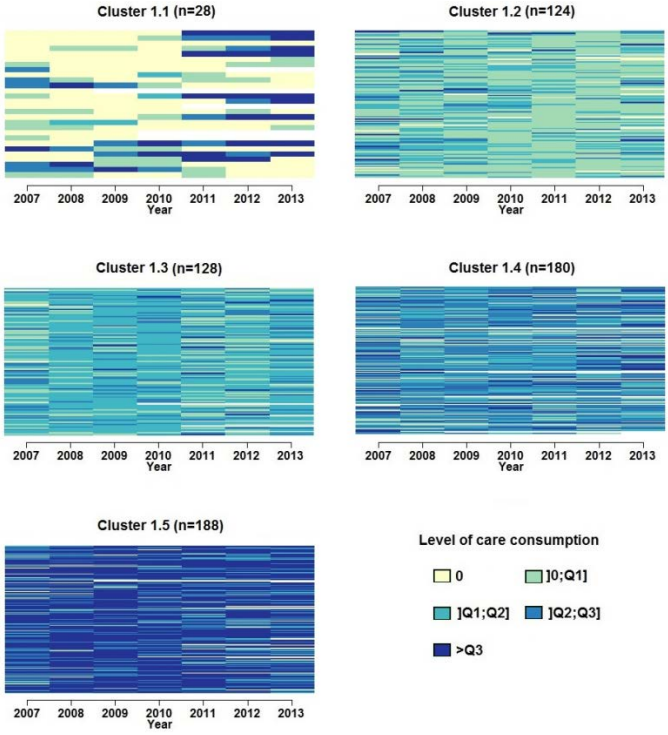
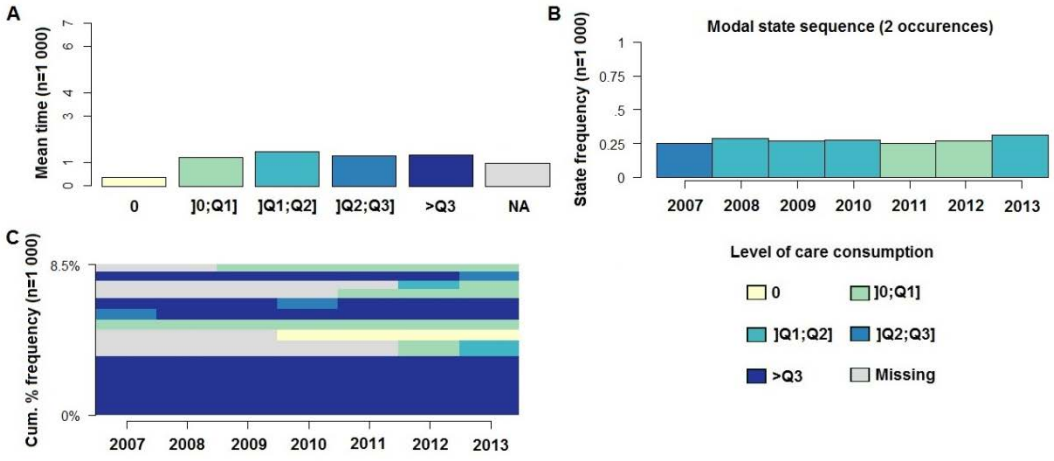


Fig.3 Index plots of clusters obtained for the care pathways of patients identified in 2007 after clustering with indel costs fixed at 0.9 and transition-based substitution costs (n=648)



Supplementary Fig.1 Graphs representing the mean duration in each state (A), the sequence composed of the modal state (i.e. the most common state at each time unit) (B), or the most frequent sequences (C) of the set of care pathways (n=1 000)