Tracking clusters of patients over time enables extracting information from medico-administrative databases

Judith Lambert, Anne-Louise Leutenegger, Anne-Sophie Jannot, Anaïs Baudot

Tracking clusters of patients over time enables extracting information from medico-administrative databases

Judith Lambert[1,2,3], Anne-Louise Leutenegger[4], Anne-Sophie Jannot[2,5,6,*] and Anaïs Baudot[3,7,8,*]

[1]Sorbonne Université, Université Paris Cité, INSERM, Centre de Recherche des Cordeliers, F-75006 Paris, France
[2]HeKA, Inria Paris, F-75015 Paris, France
[3]Aix Marseille Univ, INSERM, MMG, UMR1251, Marseille, France
[4]Université Paris Cité, INSERM, NeuroDiderot , UMR1141, 75019 Paris, France
[5]Université Paris Cité, Sorbonne Université, INSERM, Centre de Recherche des Cordeliers, F-75006 Paris, France
[6]French National Rare Disease Registry (BNDMR), Greater Paris University Hospitals (AP-HP), Paris, France
[7]CNRS, Marseille, France
[8]Barcelona Supercomputing Center, Barcelona, Spain

Corresponding author: Judith Lambert, PariSanté Campus, 10 rue d'Oradour-sur-Glane, 75015 Paris, France, judith.lambert@inserm.fr

Abstract

Context Identifying clusters (i.e., subgroups) of patients from the analysis of medico-administrative databases is particularly important to better understand disease heterogeneity. However, these databases contain different types of longitudinal variables which are measured over different follow-up periods, generating truncated data. It is therefore fundamental to develop clustering approaches that can handle this type of data.

Objective We propose here cluster-tracking approaches to identify clusters of patients from truncated longitudinal data contained in medico-administrative databases.

Material and Methods We first cluster patients at each age. We then track the identified clusters over ages to construct cluster-trajectories. We compared our novel approaches with three classical longitudinal clustering approaches by calculating the silhouette score. As a use-case, we analyzed antithrombotic drugs used from 2008 to 2018 contained in the Échantillon Généraliste des Bénéficiaires (EGB), a French national cohort.

Results Our cluster-tracking approaches allow us to identify several cluster-trajectories with clinical significance without any imputation of data. The comparison of the silhouette scores obtained with the different approaches highlights the better performances of the cluster-tracking approaches.

Conclusion The cluster-tracking approaches are a novel and efficient alternative to identify patient clusters from medico-administrative databases by taking into account their specificities.

Keywords: longitudinal clustering, cluster tracking, medico-administrative databases, patient networks

1 Introduction

The reuse of medico-administrative databases is nowadays extremely popular. Such databases are indeed increasingly available for epidemiological, clinical and healthcare research to study a large range of

*These authors contributed equally to this work.

health-related issues [1]. However, medico-administrative databases are complex and appropriate analysis methods are required [2]. First, each patient is described through a large number of variables. Analysis methods able to deal with high dimensional data are hence needed. Second, these variables are of a different nature (e.g., drug reimbursements, diagnoses, hospitalizations), and the methods need to consider heterogeneity. Finally, the variables vary over time and are measured over different follow-up periods, thereby generating truncated data when focusing on a given stage of life or disease. This time dimension is very difficult to apprehend and, overall, only few methods can deal with high dimensional truncated longitudinal data.

Among the various objectives targeted by the reuse of medico-administrative databases, the identification of clusters (i.e., subgroups) of patients is particularly significant. Indeed, given the complexity and the heterogeneity of human diseases, we have to move from a "one size fits all" paradigm towards a more personalized care and a better understanding of disease heterogeneity [3, 4]. In general, clusters of patients related to a given disease are identified using the coded diagnoses. However, in medico-administrative databases, the diagnoses are often missing due to truncated patient history. For example, if a patient had an infarction twenty years ago, the hospital stay related to this event will not be available in the database but the patient will still have treatments for secondary prevention of cardiovascular diseases. Patient history could hence be inferred from their current treatments.

To the best of our knowledge, three categories of approaches are available to cluster patients using longitudinal data. These longitudinal clustering approaches are raw-data-based, feature-based and model-based [5]. In raw-data-based approaches, classical (non-longitudinal) clustering algorithms, such as Kmeans, adapt their similarity measure to be applied to the raw longitudinal data. For instance, Kmeans adapted to raw longitudinal data has been used to identify clusters of children based on inattention and hyperactivity during elementary school [6], or to assess the relationships between fibrosis and bioclinical parameters [7]. In feature-based approaches, features are first extracted from the raw longitudinal data. These extracted features are then used as input for classical (non-longitudinal) clustering algorithms. For instance, Wang, Smith, and Hyndman extracted several features from longitudinal data in three (non-clinical) benchmark datasets [8]. They then used the extracted features as input in hierarchical clustering and in an unsupervised neural network algorithm. Although only a small number of features are used for the clustering, the identified clusters are similar to the clusters identified using all the data. Finally, model-based approaches assume that the raw longitudinal data are generated by a mixture of models and intend to extract the parameters of these models. Model-based approaches are, to the best of our knowledge, the most frequently used in biomedical research. The two prevailing model-based approaches are Growth Mixture Modeling (GMM) and Latent Class Growth Analysis (LCGA) [9]. These methods identify clusters of patients based on the common evolution of their longitudinal variables over time. GMM allows small variations around this common evolution between patients within cluster whereas LCGA assumes no variation [10]. Mora et al. applied GMM to identify clusters of women according to the magnitude and timing of depressive symptomatology from pregnancy to two years postpartum [11]. Colder et al. also used GMM to identify clusters of adolescents based on their smoking behavior over four years [12]. LGCA was used by Downie et al. to identify clusters of patients with acute low back pain from pain scores over twelve weeks [13] and by Landa et al. to identify clusters of babies at high risk for autism based on their language, motor and nonverbal cognitive functioning from 6 to 36 months [14].

However, raw-data-based, feature-based and model-based longitudinal clustering approaches have

some limitations. For instance, truncated data are not handled. Patients with truncated data must be removed or their data must be imputed. In the context of medico-administrative databases, truncated data are an inescapable issue, as patients are followed-up over a fixed period. In addition, the number of clusters must be specified *a priori*. To determine the optimal number of clusters, criteria are usually used to assess the quality of the clustering [15]. These criteria include for instance the silhouette score [16] [17] or the Davies-Bouldin criterion [18] [19]. However, the optimal number of clusters might differ depending on the criterion chosen [20]. Another limitation specific to the model-based approaches is that the majority of the studies focus on only one longitudinal variable. The joint analysis of two or three longitudinal variables is possible ([21], [22], [23], [24]), but becomes computationally challenging for more than three variables. Finally, in all three categories of approaches, each patient is assigned to only one cluster over the entire time period.

An alternative strategy for clustering patients from longitudinal data could be cluster tracking. Cluster tracking is an approach mainly used in the field of social network analysis [25]. It is a two-step strategy. In the first step, the clusters are identified at each time point. In the second step, the clusters are matched between the different time points to allow their tracking along the timeline. Clusters are identified at each time point using non-longitudinal clustering algorithms [26, 27].

Different methods can be used to identify clusters including methods such as Kmeans or network clustering algorithms. For instance, Li et al. constructed a patient network based on clinical similarity and performed a clustering approach in order to identify subtypes of type 2 diabetes [28]. Wang et al. constructed patient networks from omics data and identified clusters of cancer patients with different survival profiles [29]. Patient networks have the advantage of preserving privacy because the interactions between patients are considered rather than absolute data [30]. In addition, a large number of algorithms exist for clustering networks [31]. However, to our knowledge, current network-based approaches to identify patient clusters do not consider longitudinal data.

We propose here novel cluster-tracking approaches to identify patient clusters and trajectories from longitudinal data contained in medico-administrative databases. Our approaches starts by identifying clusters of patients at each time step. Patient clusters are identified using two clustering strategies: Kmeans directly applied to the raw data or the Markov Cluster algorithm (MCL) applied to patient networks constructed from raw data. We then track the clusters identified at the different time steps based on their sharing of patients. As a use-case, we analyzed drug reimbursements contained in the national cohort managed by the French health insurance, called the Échantillon Généraliste des Bénéficiaires (EGB). Our aim was to identify clusters of patients that could be related to given diseases using only drug reimbursements and in the absence of any coded diagnoses. We identified different trajectories of patient clusters with clinical interest. Finally, we compared these cluster-tracking approaches with three existing types of longitudinal clustering approaches, by calculating a modified silhouette score. The best modified silhouette scores were obtained with the two cluster-tracking approaches.

2 Material and methods

2.1 Cluster-tracking approach

We propose novel approaches for clustering patients from longitudinal data extracted from medico-administrative databases. These approaches start by identifying clusters of patients at each time step.

To this goal, we used two different clustering strategies: the Markov Cluster algorithm (MCL) applied to patient networks built from raw data and Kmeans applied directly on raw data. Clusters are then tracked over time steps to define cluster-trajectories.

2.1.1 Identifying clusters of patients from patient networks

The first clustering strategy used to identify clusters of patients relies on the construction of patient networks. We started by constructing a patient network for each time step. We then applied the MCL clustering algorithm on each network.

Constructing patient networks A patient network is a graph $G = (V, E)$ with V patient nodes and E edges representing interactions between patient nodes. We built a network for each time step. Each network is constructed using a similarity matrix $M_i = [m_{p_1,p_2}]^n$ where n is the number of patients, i is the time step and m_{p_1,p_2} is the similarity between patients p_1 and p_2 at the time step i. This similarity matrix is symmetrical, with $m_{p_1,p_2} = m_{p_2,p_1}$.

The similarity between patients at time step i can be computed using different similarity measures. We tested four different similarity measures: the Cosine similarity, the opposite of the normalized Euclidean distance, the Jaccard index and the generalized Jaccard index (*Supplementary section S1*).

The similarity matrices built for each time step are then filtered according to a threshold t. The goal of the filtering step is to obtain networks with a reduced number of edges [32]. The filtered matrices are next used to build patient networks. We tested different thresholds. For each threshold t, the filtered matrix M_i^t is obtained as follows:

$$M_i^t = \begin{cases} m_{p_1,p_2} & \text{if } m_{p_1,p_2} \geq t \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where a null value indicates that patients p_1 and p_2 have a similarity value below the threshold t and will thereby not be connected in the patient network. From each similarity matrix M_i^t, the associated patient network can be constructed. An edge between patients P_1 and P_2 is weighted by the value m_{p_1,p_2} of the matrix.

Reducing the number of edges may lead to disconnected nodes. Therefore, we selected the threshold t in the similarity matrices which allowed us to obtain the minimum number of isolated patient nodes in any network (*Supplementary section S2*).

Clustering patient networks We applied the Markov Cluster algorithm (MCL) [33] on the largest connected component of the patient networks. The MCL algorithm uses random walks to simulate flows on the network. The flows allow to distinguish network areas where nodes are strongly connected, which correspond to the clusters. We used the version 0.0.6.dev0 of the "markov-clustering" Python package with the default parameters.

2.1.2 Identifying clusters of patients from raw data

We described in the previous section a clustering strategy based on patient networks. We also used Kmeans as a second clustering strategy [34]. Kmeans is applied directly on raw data, for each time step.

In Kmeans, the number of clusters must be specified *a priori*. We determined the optimal number of clusters per time step by calculating the silhouette score [35]. The silhouette score assesses the clustering quality by computing the separation distance between the obtained clusters.
Let us define

$$a^i(p) = \frac{1}{|C_p^i| - 1} \sum_{j \in C_p^i, j \neq p} d(p, j), \tag{2}$$

the mean distance of patient p to their cluster C_p^i at time step i, with $|C_p^i|$ the number of patients in C_p^i and $d(p, j)$ the Euclidean distance between patients p and j belonging to C_p^i, and let

$$b^i(p) = \min_{C_z^i \neq C_p^i} \frac{1}{|C_z^i|} \sum_{z \in C_z^i} d(p, z), \tag{3}$$

be the mean distance of a patient p to their neighboring cluster C_z^i at time step i, with $|C_z^i|$ the number of patients in C_z^i and $d(p, z)$ the Euclidean distance between the patient p belonging to C_p^i and the patient z belonging to C_z^i.
We start by calculating the silhouette score for each patient at time step i as follows:

$$s^i(p) = \frac{b^i(p) - a^i(p)}{\max(a^i(p), b^i(p))}, \tag{4}$$

The silhouette score at a given time step i over all the patients is obtained as follows:

$$S^i = \frac{1}{K^i} \sum_{k=1}^{K^i} \frac{1}{|C_k|} \sum_{p \in C_k} s^i(p), \tag{5}$$

with K^i the number of clusters at time step i, $|C_k|$ the number of patients in the cluster C_k.
The silhouette score varies between -1 and 1. Values close to 1 indicate that the clusters are well-separated. Values close to 0 indicate overlapping clusters. Negative values indicate that the clusters are worse than random.

2.1.3 Tracking the clusters over time steps

In the previous step, we identified sets of clusters per time step either from patient networks with MCL or from raw data with Kmeans. We then intend to follow the clusters over the different time steps. Let C^i and C^{i+1} be two sets of clusters identified at 2 consecutive time steps, i and $i + 1$. We computed the intersection (i.e., the number of common patients) between every pair of clusters (c, c') obtained at 2 consecutive time steps:

$$Q^i(c, c') = |c \cap c'| \; \forall i, \tag{6}$$

with $c \in C^i$ and $c' \in C^{i+1}$.
Next, for each cluster $c \in C^i$, we identified the cluster from the set of clusters C^{i+1} having the greatest number of common patients as follows:

$$T_c^i = \underset{c'}{\operatorname{argmax}} \, Q^i(c, c'). \tag{7}$$

Please note that if, for the cluster c, there is more than one cluster match in T_c^i (i.e., if there is more than one cluster with the same maximum number of common patients), all the clusters are included in T_c^i.

We visualized the tracking of clusters with an alluvial plot, in which the blocks represent the clusters and the stream fields between the blocks represent the number of common patients. The height of the blocks and the thickness of the stream fields are proportional to the number of patients.

2.1.4 Identifying cluster-trajectories

We identified in the previous section sets of successive clusters. We called the sets of successive clusters cluster-trajectories. Patients in the same cluster-trajectory are considered to follow the same evolution over time for the longitudinal variables of interest.

The cluster-trajectories are visualized using a flowchart composed of blocks representing the clusters. The arrow thickness between the blocks represents the number of common patients. All clusters identified are described using the meta-information available for the patients.

2.2 Longitudinal clustering approaches

We compared the performance of the cluster-tracking approaches proposed in this work to existing state-of-the-art approaches dedicated to clustering patients using longitudinal data. The three categories of state-of-the-art longitudinal clustering approaches are raw-data-based, feature-based and model-based approaches [5, 36]. We selected three specific methods, each representative of a category of approach. All longitudinal clusters identified with these methods are described using the meta-information available for the patients.

2.2.1 Raw-data-based approach

Raw-data-based approaches work directly with longitudinal raw data [5, 36]. We selected Kml3d, an R package providing an implementation of Kmeans specifically designed for longitudinal data [37]. This package takes as input a 3-dimensional matrix $M(n, i, y)$ with n the patients, i the time step and y the set of variables characterizing the patients. The algorithm calculates the Euclidean distance between all patients (in n-dimensional space). Patients with the smallest distance are grouped in the same cluster. Importantly, the number of cluster needs to be defined *a priori*.
Kml3d cannot handle truncated data but allows imputation using different methods. We used the copy mean method (default), which imputes data using a linear interpolation and adds a variation to adapt the shape of the interpolation to the shape of the mean of the other values [38]. Patients are removed from the analysis when their number of truncated data are greater than $|I| - 2$, with I the set of time steps.

2.2.2 Feature-based approach

Raw data usually have a high dimension. The goal of the feature-based approaches is to reduce the dimensions by extracting several features characterizing the longitudinal data [5, 36]. These features can then be used as input in classic (non-longitudinal) clustering algorithms, such as Kmeans or hierarchical clustering. We extracted the most common features: mean, standard deviation, kurtosis and skewness [39]. The kurtosis and the skewness describe the shape of the distribution of longitudinal data. We

therefore obtained four features per patient and per longitudinal variable. These features were used as input in Kmeans.

2.2.3 Model-based approach

In model-based approaches, each longitudinal variable is characterized by a model or a mixture of models [5, 36]. We applied Growth Mixture Modeling (GMM), which assumes that a model with a given mean and shape is associated with each cluster [10]. Let y_p be a longitudinal variable of the patient p composed of j repeated observations and K the number of clusters, distributed with probabilities π_k with $k = 1, ..., K$, $\pi_k \in [0, 1]$ and $\sum_k \pi_k = 1$. A growth mixture modeling is defined as follows:

$$y_{p,j|k} = \beta^k_{0p} + \beta^k_{1p} \cdot i_j + \epsilon^k_{pj}, \tag{8}$$

with i_j the time step at the jth observation of the variable y, ϵ^k_{pj} the time-specific residual errors, and $(\beta^k_{0p}, \beta_{1p})$ the patient-specific coefficients.

In GMM, analyzing several variables simultaneously is computationally challenging. GMM can be applied separately for each variable, but this assumes that all longitudinal variables are independent from each other. We hence decided to use an aggregated variable $Y_p = [\sum_{v^i \in V^i_p} v^i \ \forall i \in I_p]$, with I_p the set of time steps of the patient p and V^i_p the set of longitudinal variables of the patient p at time step i. This aggregated variable allows us to apply a single GMM.
GMM calculates for every patient their posterior probability of belonging to each cluster using this aggregated variable as input. The cluster assigned to each patient is the one with the greatest posterior probability.

2.2.4 Determining the optimal number of clusters

In the raw-data-based, the feature-based and the model-based approaches, the number of clusters must be specified as a parameter *a priori*. In order to determine the optimal number of clusters, we calculated several classic clustering quality criteria (*Supplementary section S3*). In the raw-data-based and the feature-based approaches, we calculated the Calinski-Harabasz criterion [40], the Kryszczuk variant of Calinski-Harabasz criterion [41], the Genolini variant of Calinski-Harabasz criterion [37], the opposite of Ray-Turi criterion [42] and the opposite of Davies-Bouldin criterion [43]. In the model-based approach, we calculated the Akaike Information Criterion (AIC) [44] and the Bayesian Information Criterion (BIC) [45]. Furthermore, for all the approaches, we calculated a modified silhouette score as follows:

$$S = \frac{1}{|I|} \sum_{i \in I} S^i, \tag{9}$$

with S^i the silhouette score at the time step i (equation 5) and I the set of time steps. In this modified silhouette score, we calculated the silhouette score S^i at each time step rather than over the entire period. This avoids imputing truncated data.

2.3 Choice of the metric to compare the performances of the different approaches

In the cluster-tracking approaches, we used two clustering strategies: one based on network (section 2.1.1) and one based on raw data (section 2.1.2). In order to compare the clustering quality of these two clustering strategies, we calculated the modified silhouette score (equation 9). We also calculated this modified silhouette score in the three longitudinal-clustering approaches. This allowed us to compare the clustering quality of the different approaches.

We estimated the 95% confidence interval of the modified silhouette score using the percentile bootstrap method [46]. We generated 100 bootstrap samples by resampling with replacement patients present in the population of interest. In each bootstrap sample, we applied the different approaches and we calculated the modified silhouette score. We obtained the confidence interval by taking the $2.5th$ and the $97.5th$ percentile of the distribution of the modified silhouette scores.

2.4 Use-case: the Echantillon Généraliste des Bénéficiaires

We used longitudinal health data from the Echantillon Généraliste des Bénéficiaires (EGB), a French medico-administrative database. The EGB is a random sample from the French health insurance database [47]. It is representative of the French population and contains approximately 660,000 individuals followed over a period of 11 years. This study has been declared to INSERM (Institut National de la Santé et de la Recherche Médicale, `https://www.inserm.fr/`). The information provided to individuals in EGB on the possible reuse of their data and the procedures for exercising their rights comply with the legislative and regulatory provisions applicable to the processing of personal data in the SNDS (Système National des Données de Santé, `https://www.snds.gouv.fr/SNDS/Accueil`). According to French regulation, individuals in SNDS database are informed of the reuse of their data for research and can opposed to this reuse as defined by Articles 92 to 95 of Decree No. 2005-1309 of 20 October 2005 (`https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000037300884/`). As required from French regulation, EGB data can be reuse for research projects from authorized persons once the research project is declared to their institution (INSERM).

Among others, EGB contains drug reimbursements, which are longitudinal high dimensional data that can be used to identify subgroups of patients (*Figure 1*). We extracted data on drugs reimbursements between 2008 and 2018. For each patient, the date of reimbursement, the Anatomical Therapeutic Chemical (ATC) class and the name of the reimbursed drugs are indicated (see example *Table 1*). The ATC class is an international classification of drugs established by the World Health Organization (WHO) [48]. We only considered reimbursement of drugs belonging to the ATC class of antithrombotic agents (i.e., B01). We obtained 164,942 patients with such reimbursements. We further selected patients aged 60 to 70 and having had at least one drug reimbursement for two or more consecutive months. Our goal was to focus only on patients with sustained reimbursements. Our final dataset is composed of 30,111 different patients and 19 different drugs. There is a majority of men in this population, with a sex ratio (men/women) of 0.61. This is consistent with the fact that cardiovascular diseases, which accounts for the majority of antithrombotic use, is more common in men.

We also extracted data on long-term illnesses (i.e., illnesses that last at least 6 months) from the EGB. 23,063 patients out of the 30,111 patients studied experienced at least 1 long-term illness between 60 and

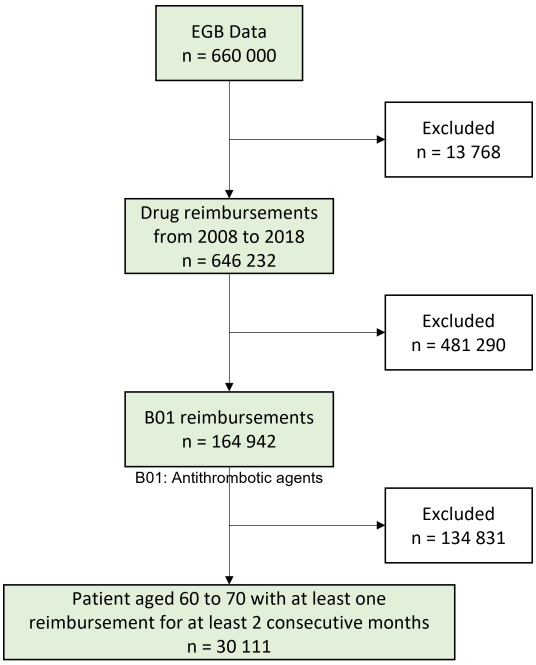

Figure 1: Extraction of longitudinal data from the EGB, considered as a use-case in this study
From the EGB medico-administrative database, we extracted antithrombotic drugs reimbursed for at least two consecutive months from 2008 to 2018 in patients ages 60 to 70. We therefore keep here only patients with sustained reimbursements. n: number of patients, B01: antithrombotic agents

70 years old. These long-term illnesses represent 865 distinct diseases. Each disease is coded with the $10th$ revision of the international statistical classification of diseases and related health problems (ICD-10 code).

Patient ID	**Reimbursement date**	**ATC class**	**Drug name**
$\boldsymbol{P_1}$	01/04/2008	M01	Ibuprofen
$\boldsymbol{P_1}$	01/12/2015	B01	Aspirin
$\boldsymbol{P_2}$	01/02/2010	N02	Tramadol
$\boldsymbol{P_3}$	01/05/2016	B01	Clopidogrel

Table 1: Example of drug reimbursements contained in the EGB
M01: Anti-inflammatory and antirheumatic products, B01: antithrombotic agents, N02: Analgesics

We decided to choose the age of the patient as time steps. Indeed, we did not have information about patients' thrombotic events nor about the initial intake of antithrombotic drugs. Choosing age as time steps is also consistent with the fact that the antithrombotic use strongly hinges on age. We hence calculated, for each patient, the number reimbursements for each drug at a given age (see example *Table 2*). We therefore obtained a table per patient age. Focusing on patients aged 60 to 70 years old, we obtained a total of 11 tables.
Importantly, we observed three types of truncated data (*Figure 2*).

In France, no more than one month's treatment can be dispensed. We therefore considered that the number of reimbursement for a drug is a good proxy for annual drug use. In the following, we suppose that when patients have reimbursements for a drug, they are exposed to that drug.

We also applied our approaches to cluster patients with primary sclerosing cholangitis contain in

pbcseq, a public database (*Supplementary section S8*). This database is a clinical trial including, among another, laboratory measurements.

Patient ID	B01_1	B01_2	B01_3
$\boldsymbol{P_1}$	0	10	5
$\boldsymbol{P_2}$	1	8	4
$\boldsymbol{P_3}$	2	6	3

Table 2: Example of total number of reimbursements that three patients aged 60 years received for three drugs
B01_1, B01_2 and B01_3 are three different drugs belonging to the ATC class B01, the antithrombotic agents

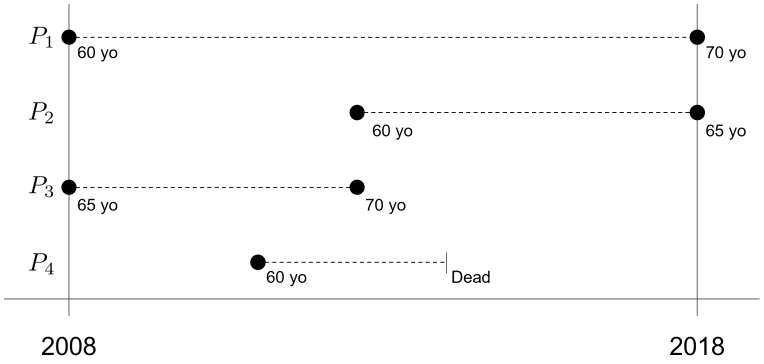

Figure 2: Example of patient follow-up in the EGB
P_1 has no truncated data because they were 60 years old in 2008 and therefore they have data for the entire period. P_2 has truncated data because they were 60 years old after 2008 and therefore they have no data before then. P_3 has truncated data because they were 70 years old before 2018 and therefore they have no data after that. P_4 has two types of truncated data because they were 60 years old after 2008 and died before 2018.

3 Results

3.1 Cluster-tracking approaches allow identifying and tracking patient clusters over ages to identify cluster-trajectories

We first apply two different clustering strategies to identify clusters of patients at each age. The first clustering strategy is applied to patient networks (Material and methods 2.1.1). The second clustering strategy is directly applied to raw data (Material and methods 2.1.2). The clusters are then tracked over ages to define cluster-trajectories.

3.1.1 Identifying cluster-trajectories with the cluster-tracking approach based on networks

The first clustering strategy used in the cluster-tracking approach relies on the construction of patient networks (Material and methods 2.1.1). Patient networks are constructed using similarity matrices. Different measures can be computed to calculate similarities between patients and construct the similarity matrices (*Supplementary section S1*). We selected the Cosine similarity because it has the greatest variance. Using this Cosine similarity, we constructed 11 similarity matrices. In each matrix, the similarities are computed between all patients of a given age (from 60 to 70 years old). For example, the 60-year-old

matrix is constructed by computing the similarities between all patients aged 60 between 2008 and 2018. Patient networks are then constructed by applying a threshold on the similarity matrices. Patients associated with a similarity higher than the threshold will be linked by an edge in the patient network. We tested different Cosine similarity thresholds and selected a threshold of 0.8. This threshold was chosen as the best trade-off to minimize the number of isolated patients while reducing the number of edges (*Supplementary section S2*). We obtained 11 patient networks (one by age, see *Table 3* and *Figure 3* for the network of patients aged 60 years old).

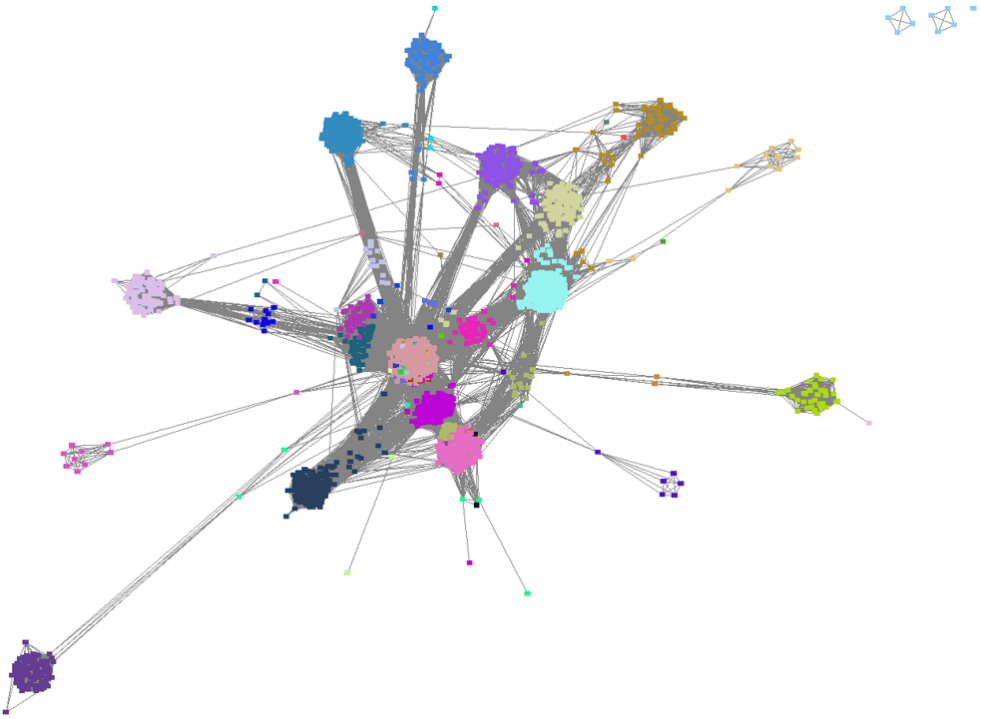

Figure 3: 60-year-old patients network

In this network, nodes represent all patients aged 60 between 2008 and 2018 and edges represent the interactions between those patients having a Cosine similarity of at least 0.8. The length of edges is inversely proportional to the Cosine similarity. Nodes of the same color belong to one of the 127 clusters identified with the Markov Cluster algorithm.

We then applied the Markov Cluster algorithm (MCL) to identify clusters of patients (Material and methods 2.1.1). The MCL algorithm is applied systematically on all the 11 patient networks, revealing different numbers of clusters per network (*Table 3*). For example, in the patient network constructed at 60 years old, 127 clusters are identified (*Figure 3*).

We next computed the number of common patients between clusters identified at consecutive ages (Material and methods 2.1.3). This allows tracking the evolution of the clusters over consecutive ages (*Figure 4*) and identifying cluster-trajectories. We identified 12 cluster-trajectories composed of clusters with at least 100 patients (*Supplementary section S4*). We described the clusters that compose these trajectories with the number of patients, the sex ratio, the two most frequently reimbursed drugs and the two most frequent long-term illnesses. Most of the 12 identified trajectories are composed of clusters with a majority of men. This is explained by the presence of a majority of men in our study population (i.e., 30,111 patients). Indeed, the sex ratio of this population is 0.61.

We next focused on the 3 cluster-trajectories (A,B and C) with the largest number of patients (*Figure*

5 and *Supplementary section S4*). The trajectory A is the one with the largest number of patients. By analyzing clusters at all ages of this trajectory, we observed that all patients used aspirin. Furthermore, more than half of the patients present in any cluster of the trajectory A are also present in the following cluster. For instance, among the 4238 patients of the cluster 60.1 identified at age 60, 3 209 (i.e., 76 %) are present in the cluster 61.1 of age 61. Thus, for the majority of the patients, aspirin is used for at least two consecutive years. In addition, at 63 and 64 years old, two clusters are observed in the trajectory A. The first cluster (63.1 and 64.1) is associated with aspirin use only and the second cluster (63.14 and 64.11) is associated with enoxaparin use in addition to aspirin. These two clusters merge into the same cluster at the following age (64.1 and 65.1) in which only aspirin is used. This implies that, when enoxaparin is used in addition to aspirin, the majority of the patients switch to aspirin-only use the following year. The most frequent long-term illnesses observed in clusters that compose this trajectory is diabetes (ICD-10 code E11). This diagnosis is also observed in all the 12 trajectories identified. The other long-term illness observed in the trajectory A is chronic ischemic heart disease (ICD-10 code I25).

The trajectory B is composed of clusters in which clopidogrel, an antiplatelet drug, is used by all patients. Two clusters are systematically observed at each age. For example at age 60, in the first cluster 60.2, clopidogrel is the only drug used. In the second cluster 60.8, aspirin is used in addition to clopidogrel. These two clusters merge into the same cluster 61.2 at the following age in which clopidogrel is the only drug used. Hence, we can observe that when aspirin is used in addition to clopidogrel, the majority of the patients switch to clopidogrel-only use the following year. The most frequent long-term illness in addition to diabetes is peripheral arterial disease (ICD-10 code I702).

The trajectory C is composed of clusters of patients who use fluindione; about 12 % of the patients also use enoxaparin. More than half of the patients present in any cluster of the trajectory C are also present in a cluster of the following year. For instance, among the 679 patients present in the cluster 60.3 identified at age 60, 503 (i.e., 74 %) are present in the cluster 61.3 identified at the age 61. Thus, we can conclude that, for the majority of the patients, fluindione is used for at least two consecutive years. The most frequent long-term illness in this trajectory is atrial fibrillation (ICD-10 code I48).

The same interpretations were carried out for the 9 remaining cluster-trajectories (*Supplementary section S4*). In each cluster that compose these trajectories, we always observe a drug used by all patients (i.e., predominant drug). Most of the time, more than half of the patients present in the clusters of these trajectories are also present in the following-age clusters. Thus, the predominant drugs are usually used for at least two consecutive years. However, this is not the case in the cluster-trajectory D. In this trajectory, two types of clusters are usually observed at each age. The first cluster contains patients who all used enoxaparin and the second cluster contains patients who all used tinzaparin. These two clusters systematically merge into the cluster 0 at the following age (e.g., cluster 61.0 at age 60). The cluster 0 is composed of patients with no antithrombotic use. Thus, the majority of patients with enoxaparin or tinzaparin use in this trajectory no longer use antithrombotics at the following year. This cluster-trajectory D is also the only one with clusters composed of a majority of women (i.e., sex ratio about 0.40). Associated comorbidities are scarce, with the most frequent long-term illnesses being cancers (ICD-10 codes C50, C34, C18).

Age	Number of nodes	Number of edges (10^7)	Number of clusters
60	8268	1.25	127
61	8884	1.46	144
62	9555	1.70	162
63	10042	1.87	149
64	10466	2.03	168
65	10761	2.15	165
66	11097	2.27	150
67	11392	2.37	168
68	11492	2.43	207
69	11664	2.45	205
70	11687	2.48	220

Table 3: Number of nodes, edges and clusters in 60 to 70 years old patient networks

3.1.2 Identifying cluster-trajectories with the cluster-tracking approach based on raw data

In the previous section (3.1.1), we identified cluster-trajectories using a network-based cluster-tracking approach. We also implemented a cluster-tracking approach using Kmeans applied to raw data (Material and methods 2.1.2). In this second strategy, we applied a Kmeans per patient age, from 60 to 70 years old.

In Kmeans, the number of clusters must be specified *a priori*. We calculated the silhouette score and identified an optimal number of clusters at each patient age (*Supplementary section S5*). The optimal number of clusters was between 6 and 8. We then tracked the clusters identified by Kmeans over ages (Material and methods 2.1.3). We identified 9 cluster-trajectories composed of clusters with at least 100 patients (*Supplementary section S6*). We described these trajectories with the number of patients, the sex ratio, the two most frequently reimbursed drugs and the two most frequent long-term illnesses. We observed that all trajectories are composed of a majority of men. This is explained by the presence of a majority of men in our study population (i.e., 30,111 patients).

For the sake of simplicity, we next focused on three cluster-trajectories (A,B and C). We represented them from 60 to 65 years old (*Figure 6*). The trajectory A is the one with the largest number of patients. Aspirin is used by all patients in the clusters that compose this trajectory. In all the clusters of the trajectory B, clopidogrel is used by all patients. In all the clusters of the trajectory C, fluindione is used by all patients and enoxaparin is used by about 12 % of patients. In addition, more than half of the patients present in any cluster of these three trajectories are also present in the following-age clusters. Thus, we can conclude that, for the majority of the patients, aspirin, clopidogrel and fluindione are used for at least two consecutive years in the trajectories A, B, and C, respectively. As in the network-based cluster-tracking approach, diabetes (ICD-10 code E11) is one of the most frequent long-term illnesses observed in clusters of all identified trajectories. The other long-term illness observed in the trajectory A is chronic ischemic heart disease (ICD-10 code I25). In trajectory B, the most frequent long-term illness in addition to diabetes in the clusters of age 60 (60.4) and 61 (61.3) is peripheral arterial disease (ICD-10 code I702). In the clusters identified from 62 years old, the most frequent long-term illness is chronic ischemic heart disease (ICD-10 code I25) In trajectory C, the most frequent long-term illness is atrial fibrillation (ICD-10 code I48).

These same descriptions were carried out for the 6 other cluster-trajectories (*Supplementary section*

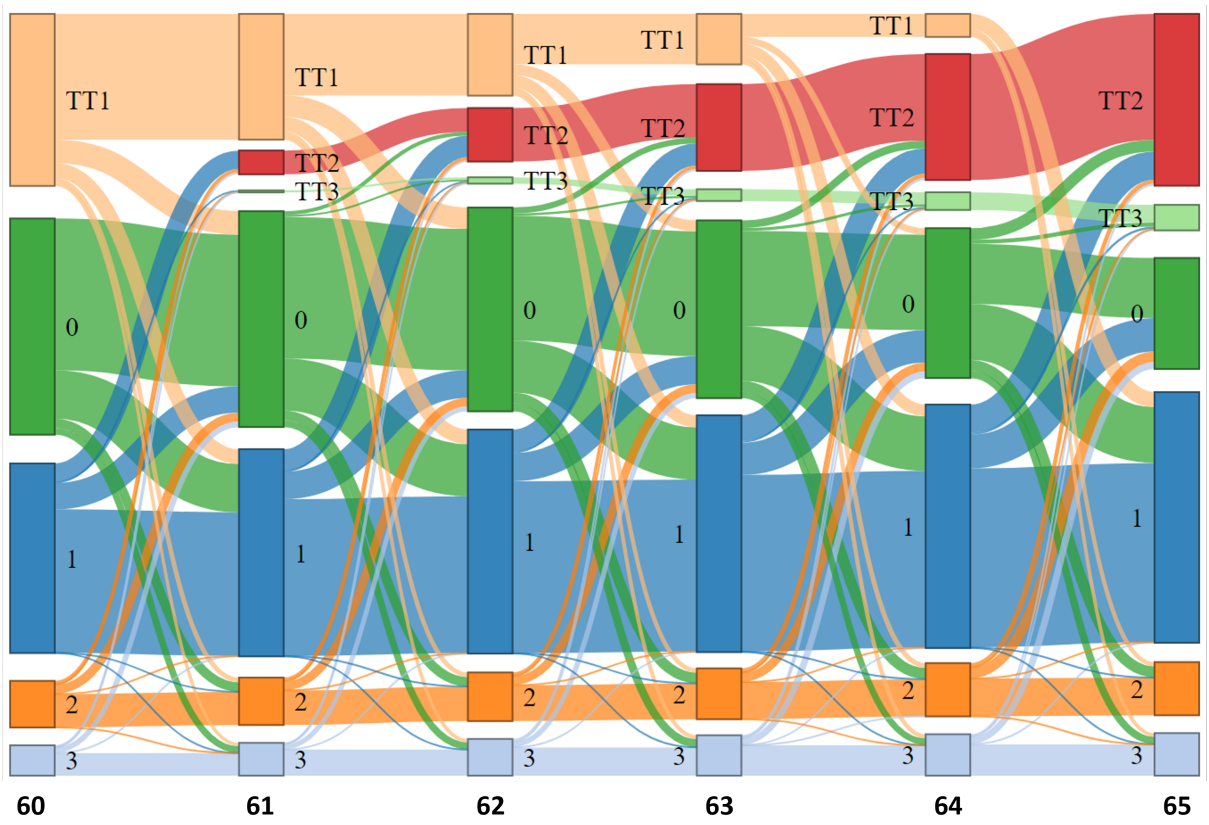

Figure 4: Tracking of clusters identified from patient networks

The alluvial plot represents the tracking of the clusters identified from 60 to 65 years old. The clusters were identified at each age based on patient networks with the MCL algorithm. At each age, the color blocks represent the different clusters of patients. Stream fields between blocks represent the number of common patients between clusters of consecutive ages. The height of the blocks and the thickness of the stream fields are proportional to the number of patients. At each age, only the clusters containing more than 500 patients are represented (corresponding to blocks 1 to 3). The blocks 0 correspond to the clusters of patients with no antithrombotic use. The three blocks TT1 (Truncated Type 1), TT2 (Truncated Type 2) and TT3 (Truncated Type 3) are the clusters of patients with truncated data. TT1 contains patients aged 70 before 2018; TT2 contains patients aged 60 after 2008 and TT3 contains patients who have died before 2018 (*Figure 2*).

S6). In each cluster that compose these trajectories, we always observe a predominant drug used by all patients. Hence, we can conclude that the predominant drug is used for at least two consecutive years. This is not the case in the cluster-trajectories D and F. In the trajectory D, several clusters merge into the cluster 0 (e.g., cluster 61.0 at age 60), which is composed of patients with no antithrombotic use. Thus, most of the patients in this trajectory no longer use antithrombotics at the following year. Contrarily to what we previously observed in the network-based cluster-tracking approach, this trajectory D is not composed of a majority of women (i.e., sex ratio about 0.53). In the trajectory F, combinations (i.e., combination of two platelet aggregation inhibitors) are used to all patients in clusters identified from 61 to 67 years old. Then aspirin is used by about 60% of patients in clusters identified from 68 years old.

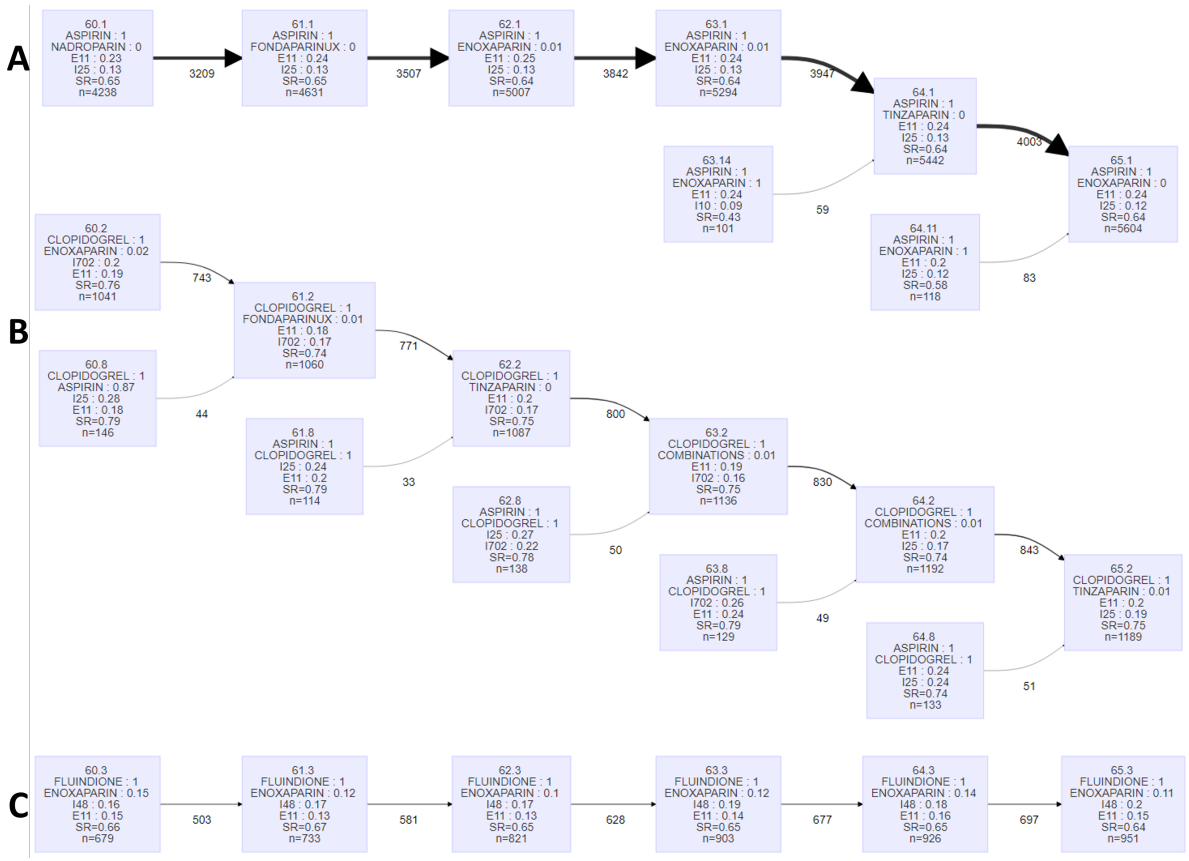

Figure 5: Subset of patient cluster-trajectories identified with the cluster-tracking approach based on network

We represented 3 cluster-trajectories (A,B and C) out of the 12 identified. We represented them from 60 to 65 years old. In these 3 cluster-trajectories, each block represents a cluster. Each cluster is named as follows: "x.y", with x the age at which it was identified and y its cluster number in the alluvial plot (*Figure 4*). The clusters are characterized by the two most frequently reimbursed drugs (name, percentage of patients receiving the drug), the two most frequent long-term illnesses (ICD-10 code, percentage of patients), the sex ratio (SR) and the total number of patients (n). The number under arrows is the number of common patients between the two blocks. The arrow thickness is proportional to this number. Combinations: combination of two platelet aggregation inhibitors. ICD-10 code E11: type 2 diabetes mellitus, I25: chronic ischemic heart disease, I10: essential primary hypertension, I702: atherosclerosis of arteries of extremities, I48: atrial fibrillation.

3.2 Comparing the two clustering strategies used in the cluster-tracking approaches

We identified the cluster-trajectories with the cluster-tracking approaches using two different clustering strategies: one based on the construction of patient networks by applying the MCL algorithm and one based on raw data by applying Kmeans. We aimed to compare the performances of these two clustering strategies.

We observed that the trajectories A in the two cluster-tracking approaches are composed of clusters having a similar description (*Supplementary sections S4 and S6*). Indeed, aspirin is used by all the patients and the two most frequent long-term illnesses are the same in all the clusters. We also observed a similar description between the clusters of the trajectories C and E of the two cluster-tracking approaches. The clusters of the two trajectories G also have a similar description, but the two trajectories do not begin

at the same age. The first cluster is identified at 60 years old with the network-based cluster-tracking approach and at 64 years old with the raw-data-based cluster-tracking approach. The two trajectories H also begin at different ages. In both cases, the cluster-trajectories identified with the network-based cluster-tracking approach start at earlier ages than the cluster-trajectories identified with the raw-data-based cluster-tracking approach.

We calculated the modified silhouette score (S) and its 95% confidence interval to assess clustering quality in the two cluster-tracking approaches (Material and methods 2.3). We obtained $S = 0.50$ ([0.46 ; 0.55]) with the network-based cluster-tracking approach and $S = 0.57$ ([0.53 ; 0.58]) with the raw-data-based cluster-tracking approach (*Table 4 B.*). *A priori*, the cluster-tracking approach seems to be more efficient using a raw-data-based than a network-based strategy. But we can observe that the confidence intervals of the modified silhouette scores obtained with the network-based and raw-based clustering approaches overlap.

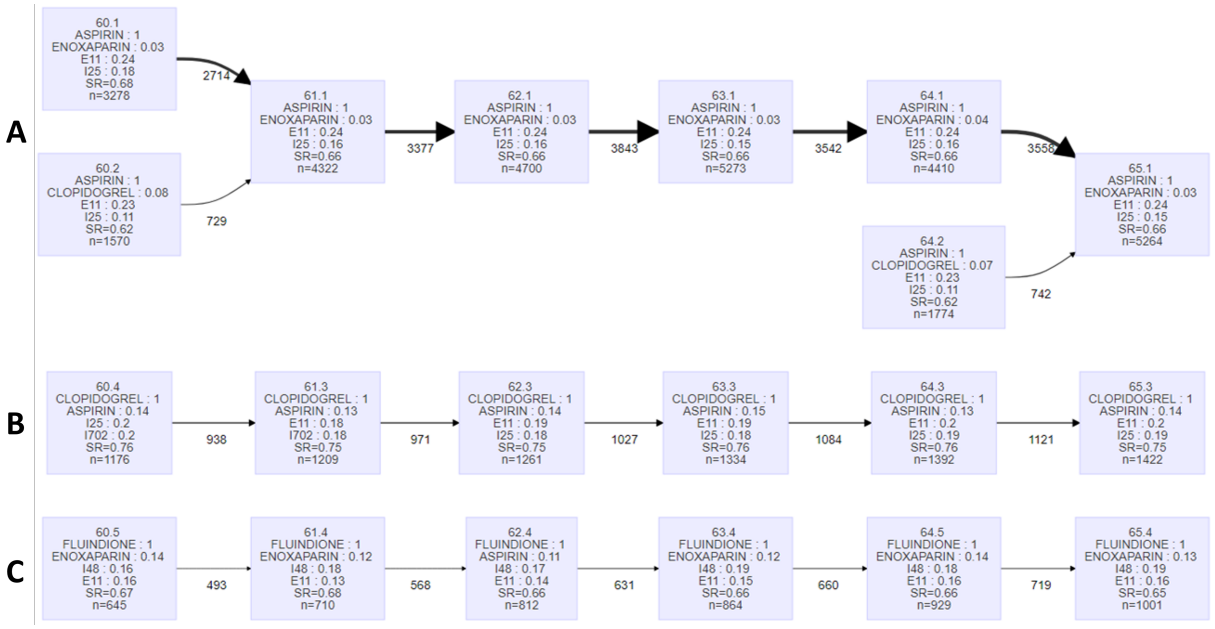

Figure 6: Subset of patient cluster-trajectories identified with the raw-data-based cluster-tracking approach

We represented 3 cluster-trajectories out (A,B and C) of the 9 identified. We represented them from 60 to 65 years old. In these 3 trajectories, each block represents a cluster. Each cluster is named as follows: "x.y", with x the age at which it was identified and y the number of the cluster. The clusters are characterized by the two most frequently reimbursed drugs (name, percentage of patients), the two most frequent long-term illnesses (ICD-10 code, percentage of patients), the sex ratio (SR) and the number of patients (n). The number under arrows is the number of patients in common between the two blocks. The arrow thickness is proportional to this number. ICD-10 code E11: type 2 diabetes mellitus, I25: chronic ischemic heart disease, I702: atherosclerosis of arteries of extremities, I48: atrial fibrillation.

3.3 Comparing the cluster-tracking approach with the longitudinal-clustering approaches

We compared the performance of the cluster-tracking approaches based on network and raw-data with three methods representative of the three types of longitudinal clustering approaches, namely raw-data-based, feature-based and model-based approaches (Material and methods 2.2). We used the same longi-

tudinal data extracted from EGB in patients aged from 60 to 70 years old in all the approaches.

3.3.1 Choosing the optimal number of clusters

In the three longitudinal-clustering approaches, the number of clusters need to be specified *a priori*. In order to select an optimal number of clusters, we calculated several classic clustering quality criteria (Material and methods 2.2.4). These criteria however do not point to clear optimums (*Supplementary section S3*). Hence, we next tried to use the modified silhouette score. We also failed to find a clear optimum with this approach. Indeed, the greatest silhouette scores (i.e., global maximum) was obtained for the smallest number of clusters (*Supplementary section S3*). We therefore decided to specify the number of clusters as 12 clusters. This number corresponds to the number of cluster-trajectories identified with the network-based cluster-tracking approach.

3.3.2 Identifying clusters with the raw-data-based longitudinal-clustering approach

We applied Kml3d [37], the selected raw-data-based longitudinal clustering approach (Material and methods 2.2.1) to the longitudinal data extracted from the EGB. First, 1737 patients are removed by the Kml3d algorithm because they have more than 9 truncated data (which is the limit with 11 different ages). We applied the Kml3d algorithm with 12 clusters as parameter and we described all the identified longitudinal-clusters with the number of patients, the sex ratio, the two most frequently reimbursed drugs and the two most frequent long-term illnesses.

Among the 12 longitudinal-clusters identified by Kml3d, 10 are composed of at least 100 patients (*Table 4 A.*). At least one of the two most frequently reimbursed drugs is used by more than 60 % of patients. For instance, aspirin, clopidogrel, combinations, warfarin and ticlopidine are used by all patients in longitudinal-clusters B, C, F, H and L, respectively. Each longitudinal-cluster identified is therefore characterized by a drug that is predominantly used by patients. More than 20 % of patients have diabetes (ICD-10 code E11) in all the longitudinal-clusters except in the longitudinal-cluster G. Atrial fibrillation (ICD-10 code I48) is one of the two most frequent long-term illnesses in longitudinal-clusters D, E, H, I and K. In these clusters, at least 70 % of patients use vitamin K antagonist (such as fluindione, warfarin or acenocoumarol) or non-vitamin K antagonist oral anticoagulants (such as rivaroxaban or apixaban). Chronic ischemic heart disease (ICD-10 code I25) is always observed in the longitudinal-clusters when aspirin is one of the two most frequently reimbursed drugs.

Our goal is to compare the 12 longitudinal-clusters obtained with the raw-data-based longitudinal clustering approach with the cluster-trajectories identified with the cluster-tracking approaches. At least one of the two most frequently reimbursed drugs is used by more than 60 % of patients in all the clusters that compose the cluster-trajectories (*Supplementary sections S4* and S6) and in all the longitudinal-clusters (*Table 4 A.*). This is not the case in the raw-data-based-cluster-trajectory D where aspirin is used by about 38 % of patients and enoxaparin is used by about 16 % of patients. Therefore, the majority of cluster-trajectories and longitudinal-clusters are characterized by a predominantly used drug. These trajectories and longitudinal-clusters are composed of a majority of men except in the network-based-cluster-trajectory D where the sex ratio is about 0.40. Breast cancer (ICD-10 code C50) is usually one of the two most frequent long-term illnesses in the clusters that compose the network-based-cluster-trajectory D. Several cluster-trajectories and longitudinal-clusters have a common drug description. For instance, aspirin and enoxaparin are both used in the longitudinal-cluster B and in the two cluster-trajectories A of

the cluster-tracking approaches. The two most frequent long-term illnesses are also the same. Conversely, the raw-data-based longitudinal clustering approach is the only one to have identified three longitudinal-clusters characterized by use of ticagrelor-aspirin, prasugrel-aspirin and ticlopidine-aspirin (G, J and L respectively in *Table 4 A.*). Similarly, the network-based cluster-tracking approach is the only one to have identified cluster-trajectories characterized by use of enoxaparin-tinzaparin, aspirin-fluindione and dabigatran-enoxaparin (D, J and L respectively in *Supplementary section S4*). Therefore, additional information are given with the raw-data-based longitudinal clustering approach and the network-based cluster-tracking approach compared to the raw-data-based cluster-tracking approach.

Furthermore, we calculated the modified silhouette score (S) in the raw-data-based longitudinal clustering approach and in the cluster-tracking approaches to compare the clustering quality (Material and methods 2.3). We obtained $S = 0.27$ for the raw-data-based longitudinal clustering approach, $S = 0.50$ for the network-based cluster-tracking approach and $S = 0.57$ for the raw-data-based cluster-tracking approach (*Table 4 B.*). The 95% confidence intervals of the two strategies of cluster-tracking approach overlap (*Table 4 B.*). Overall, we obtained a better clustering quality with the cluster-tracking approaches compared to the raw-data-based longitudinal clustering approach.

3.3.3 Identifying clusters with the feature-based longitudinal-clustering approach

We extracted 4 standard features from the the antithrombotic drug use contained in the EGB: the mean, the standard deviation, the kurtosis and the skewness (Material and methods 2.2.2). We therefore obtained a total of 76 features per patient (i.e., 4 features extracted over the 19 antithrombotic drugs). We then used these features as input in Kmeans. Here, the Kmeans clustering is applied over all the ages jointly. As for the raw-data-based longitudinal clustering approach, we applied the Kmeans clustering selecting 12 clusters as parameter. We described the identified longitudinal-clusters with the number of patients, the sex ratio, the two most frequently reimbursed drugs and the two most frequent long-term illnesses.

The 12 longitudinal-clusters identified with the feature-based longitudinal clustering approach are all composed of at least 100 patients (*Table 4 A.*). One of the two most used drugs is always used by all patients except in the cluster B. In this cluster, aspirin is used by 41 % of the patients and enoxaparin is used by 28 % of the patients. The majority of the identified longitudinal-clusters is therefore characterized by a predominantly used drug. At least 15 % of patients have diabetes (ICD-10 code E11) in all the clusters. Chronic ischemic heart disease (ICD-10 code I25) is always observed in the clusters where aspirin is one of the two most frequently reimbursed drugs.

We compared the 12 longitudinal-clusters obtained in the feature-based longitudinal clustering approach with the cluster-trajectories identified in the cluster-tracking approaches (*Supplementary sections S4* and S6). We observe that the longitudinal-clusters A and D have a common drug and long-term illness description (*Table 4 A.*). Indeed, aspirin and enoxaparin are both used by a similar proportion of patients and the two most frequent long-term illnesses are the same (i.e., ICD-10 codes E11 and I25). This type of redundant information is not observed in the cluster-trajectories identified with the two cluster-tracking approaches.

We then calculated the modified silhouette score (S) in the feature-based longitudinal clustering approach to compare the clustering quality with the other clustering approaches (Material and methods 2.3). We obtained $S = 0.20$ for the feature-based longitudinal clustering approach (*Table 4 B.*). This

score indicates that patients are less well assigned in clusters with the feature-based longitudinal clustering approach than with the cluster-tracking approach and with the raw-data-based longitudinal clustering approach. The clustering quality is therefore better with the cluster-tracking approaches.

3.3.4 Identifying clusters with the model-based longitudinal-clustering approach

The model-based approach that we applied to the antithrombotic drug use is GMM (Material and methods 2.2.3). We used an aggregated variable with this algorithm because the simultaneous analysis of several variables is computationally challenging [49]. This aggregated variable is calculated, for each patient, as the total number of drugs used at a given age. As before, we applied GMM selecting 12 clusters as parameter. We described the identified longitudinal-clusters with the number of patients, the sex ratio, the two most frequently reimbursed drugs and the two most frequent long-term illnesses.

The GMM algorithm assigns patients to the cluster for which they have the greatest posterior probability of belonging. Although we chose 12 clusters as parameter, none of the patients had a greatest posterior probability of belonging to three out of the 12 selected clusters. Therefore, only 9 longitudinal-clusters were identified.

The longitudinal-clusters A to G are composed of more than 100 patients (*Table 4 A.*). The two remaining clusters are composed of less than 20 patients. In the 9 longitudinal-clusters, we observed that aspirin is used by more than 50 % of patients. All these longitudinal-clusters are therefore characterized by the same predominantly used drug. Diabetes (ICD-10 code E11) is always one of the two most frequent long-term illnesses except in longitudinal-cluster I. The longitudinal-cluster I is very small with only two patients. One of the patients has prostate cancer (ICD-10 code C61) and the other has fibrosis and cirrhosis of liver (ICD-10 code K74).

We compared the 9 longitudinal-clusters with the cluster-trajectories identified with the cluster-tracking approaches (*Supplementary sections S4* and S6). The longitudinal-clusters are highly different compared to the cluster-trajectories. Indeed, aspirin is used by a majority of patients in all these longitudinal-clusters, which is not the case in the cluster-trajectories. Furthermore, the diversity of the two most frequently reimbursed drugs is lower in the longitudinal-clusters since only aspirin, clopidogrel, enoxaparin, fluindione or fondaparinux are observed. In the cluster-trajectories, other drugs such as warfarin, combinations or rivaroxaban are additionally observed. The model-based longitudinal clustering approach therefore identified longitudinal-clusters where patients are more heterogeneous compared to the cluster-tracking approach.

We then calculated the modified silhouette score (S) in the model-based longitudinal clustering approach to compare the clustering quality with the other clustering approaches (Material and methods 2.3). We obtained $S = -0.33$ for the model-based longitudinal clustering approach (*Table 4 B.*). This negative score indicates that the clusters are worse than random. The model-based longitudinal clustering approach therefore fails to identify patient clusters. Among all the analyzed approaches, the best clustering quality is obtained with the cluster-tracking approaches.

A.

	Raw-data-based longitudinal-clustering				Feature-based longitudinal-clustering				Model-based longitudinal-clustering			
	n	**SR**	**Top 2 drugs (%)**	**Top 2 diseases (%)**	**n**	**SR**	**Top 2 drugs (%)**	**Top 2 diseases (%)**	**n**	**SR**	**Top 2 drugs (%)**	**Top 2 diseases (%)**
A	12550	0.53	Aspirin (65) Enoxaparin (22)	E11 (23) I25 (7)	11510	0.62	Aspirin (100) Enoxaparin (11)	E11 (32) I25 (17)	14461	0.65	Aspirin (76) Clopidogrel (27)	E11 (28) I25 (18)
B	8665	0.64	Aspirin (100) Enoxaparin (15)	E11 (32) I25 (21)	7484	0.51	Aspirin (41) Enoxaparin (28)	E11 (17) I48 (6)	6822	0.50	Aspirin (52) Enoxaparin (23)	E11 (19) I10 (6)
C	2937	0.75	Clopidogrel (100) Aspirin (60)	E11 (29) I25 (28)	2827	0.73	Clopidogrel (100) Aspirin (44)	E11 (27) I25 (24)	2481	0.77	Aspirin (92) Clopidogrel (66)	I25 (44) E11 (29)
D	1794	0.65	Fluindione (99) Enoxaparin (43)	I48 (24) E11 (22)	2460	0.62	Aspirin (100) Enoxaparin (20)	E11 (31) I25 (15)	2198	0.59	Aspirin (74) Enoxaparin (18)	E11 (29) I10 (10)
E	1013	0.65	Rivaroxaban (70) Aspirin (41)	I48 (37) E11 (23)	2050	0.60	Fluindione (100) Enoxaparin (38)	I48 (24) E11 (19)	2033	0.55	Aspirin (70) Fluindione (17)	E11 (25) I10 (8)
F	402	0.81	Combinations (100) Aspirin (79)	I25 (45) E11 (28)	1114	0.74	Clopidogrel (100) Aspirin (89)	I25 (41) E11 (30)	1296	0.61	Aspirin (82) Enoxaparin (22)	E11 (30) I25 (13)
G	345	0.77	Ticagrelor (98) Aspirin (97)	I25 (46) I21 (33)	615	0.77	Aspirin (100) Clopidogrel (100)	I25 (37) E11 (31)	803	0.58	Aspirin (78) Fluindione (18)	E11 (31) I10 (12)
H	243	0.62	Warfarin (100) Enoxaparin (46)	E11 (23) I48 (19)	576	0.62	Rivaroxaban (100) Aspirin (16)	I48 (32) E11 (16)	15	0.80	Aspirin (100) Clopidogrel (87)	I25 (67) E11 (13)
I	233	0.64	Acenocoumarol (99) Enoxaparin (42)	E11 (23) I48 (18)	509	0.80	Combinations (100) Aspirin (97)	I25 (51) E11 (33)	2	1.00	Aspirin (100) Fondaparinux (50)	C61 (50) K74 (50)
J	106	0.82	Prasugrel (96) Aspirin (93)	I25 (62) E11 (28)	454	0.70	Fluindione (100) Aspirin (96)	E11 (28) I48 (22)				
K	73	0.66	Apixaban (82) Aspirin (33)	I48 (30) E11 (18)	410	0.63	Rivaroxaban (100) Aspirin (60)	I48 (34) E11 (23)				
L	13	0.85	Ticlopidine (100) Aspirin (54)	E11 (23) C34 (15)	102	0.62	Warfarin (100) Aspirin (61)	E11 (26) I25 (17)				

B.

Network-based cluster-tracking	Raw-data-based cluster-tracking	Raw-data-based longitudinal-clustering	Feature-based longitudinal-clustering	Model-based longitudinal-clustering
0.50 [0.46 ; 0.55]	0.57 [0.53 ; 0.58]	0.27	0.20 [-0.02 ; 0.20]	-0.33 [-0.26 ; 0.01]

Table 4: Longitudinal-clusters identified with the three longitudinal clustering approaches and comparison with the cluster-tracking approaches

A. n: number of patients, SR: sex ratio (percentage of men), Top 2 drugs: the two most frequently reimbursed drugs with the percentage of patients, Top 2 diseases: the two most frequent long-term illnesses (ICD-10 code) with the percentage of patients. In all approaches, the identified longitudinal-clusters are ranked from the largest to the smallest. Combinations: combination of two platelet aggregation inhibitors. ICD-10 codes C34: malignant neoplasm of bronchus and lung, C50: malignant neoplasm of breast, C61: malignant neoplasm of prostate, E11: type 2 diabetes mellitus, I10: essential primary hypertension, I21: acute myocardial infarction, I25: chronic ischemic heart disease, I702: atherosclerosis of arteries of extremities, I48: atrial fibrillation, K74: fibrosis and cirrhosis of liver.

B. silhouette scores calculated in the different approaches and their 95% confidence intervals.

4 Discussion

We proposed here novel approaches based on cluster-tracking, with the objective of clustering patients using longitudinal data extracted from medico-administrative databases. We applied these new approaches to the analysis of antithrombotic drugs extracted from the Echantillon Généraliste des Bénéficiaires (EGB). We extracted the data from 2008 to 2018 and focused on patients aged from 60 to 70 years old. We aimed to identify clusters of patients that could be related to given diseases using only drug reimbursements and in the absence of any coded diagnoses. We showed that cluster-tracking approaches are efficient to identify patient trajectories from medico-administrative databases. They are able to consider the longitudinal, multidimensional and truncated nature of data. We were able to identify clusters of patients

related to given diseases based only on drug reimbursements. We compared these new cluster-tracking approaches with three classical longitudinal clustering approaches using a modified silhouette score. We showed that the cluster-tracking approaches had a higher performance than the classical approaches.

We here applied all the approaches using age as time steps. However, it is to note that different data types can be used as input of our approaches. Depending on those input data, different time steps can be chosen. For example, we applied our two cluster-tracking approaches to the pbcseq database [50] using patient visits as time steps. This allowed us to cluster patients with primary sclerosing cholangitis based on their laboratory measurements (*Supplementary section S8*).

We identified 12 and 9 cluster-trajectories with the network-based and raw-data-based cluster-tracking approaches, respectively. We described the clusters that compose the cluster-trajectories with their number of patients, sex ratio, the two most frequently reimbursed drugs and the two most frequent long-term illnesses. Of note, for both approaches, the top three largest cluster-trajectories had similar characteristics. The trajectories with the highest number of patients identified with the two cluster-tracking approaches (trajectories A) are composed of patients with aspirin use and chronic ischemic heart disease. Antithrombotic therapy is a key part of secondary prevention in patients with chronic ischemic heart disease and patients with this illness are considered for long-term aspirin treatment [51]. The trajectories B identified with the two cluster-tracking approaches are composed of patients with clopidogrel use and coded arteriopathies as long-term illnesses (i.e., peripheral arterial disease and chronic ischemic heart disease). This is in accordance with clopidogrel being the preferred antiplatelet drug indicated in patients with arteriopathies that are symptomatic or have undergone revascularization [52]. The trajectory B identified with the network-based cluster-tracking approach also shows patients using aspirin with clopidogrel and switching to the use of clopidogrel-only the following year. This is in accordance with the fact that after myocardial infarction and percutaneous coronary intervention, a switch to mono-therapy is recommended after one year of dual antiplatelet [53]. The two trajectories C, the third largest trajectories identified with the two cluster-tracking approaches, are composed of patients with fluindione use and coded atrial fibrillation. Fluindione, which is a vitamin K antagonist, has been shown to strongly reduce stroke in patients with atrial fibrillation [54]. Furthermore, in the trajectory C identified by the network-based cluster-tracking approach, we observed a switch of drugs from age 67. Recently, non-vitamin K antagonist oral anticoagulants (e.g., apixaban and rivaroxaban) have been recommended in replacement of vitamin K antagonists [55]. Because non-vitamin K antagonist oral anticoagulants are more convenient to use, the switch of drugs observed from age 67 with the network-based cluster-tracking approach is consistent. The two identified trajectories D are composed of patients using low molecular weight heparin (i.e., enoxaparin or tinzaparin) over a short period of time. Indeed, these patients do not use antithrombotics the following year. We hypothesize that these trajectories captured patients having an acute venous thromboembolism event. However, the trajectory D identified with the network-based cluster-tracking approach was the only one composed of patients who were mostly women with cancers. There is a known significant increase of thromboembolism event requiring low molecular weight heparin in these patients [56]. Moreover, it is well-known that women have a higher risk of thromboembolism event than men [57]. The trajectory F identified with the raw-data-based cluster-tracking approach was the only one with patients first using combination of two antithrombotic drugs and then aspirin at older age. As hemorrhage risk increases with age, patients at older age switch to only one platelet aggregation inhibitor [53]. As a side note, regarding long-term illnesses, diabetes was among the two most frequent

long-term illnesses in all the cluster-trajectories. No specific antithrombotic drugs are recommended for patients suffering from diabetes. However, diabetes increases cardiovascular risk and therefore many patients with antithrombotic drugs have diabetes [58].

We compared these new cluster-tracking approaches with three classical longitudinal clustering approaches. The better modified silhouette score was obtained with the cluster-tracking approaches. This higher performance might arise from a better usage of the available information. Indeed, clustering per age allows us to take into account a maximum number of patients: as the clustering is performed by age, patient follow-ups over the entire period are not required and missing data can be handled. Contrarily, classical longitudinal clustering approaches require patient follow-ups over the entire period. Longitudinal clustering approaches hence either impute data or exclude patients with truncated data. Our new cluster-tracking approaches are therefore less sensitive to small sample sizes. However, large sample sizes increase computation time for all the approaches (*supplementary Table S1*). Another interesting feature of the cluster-tracking approaches is that patients can switch clusters as their age progresses. A patient can therefore belong to several cluster-trajectories. This allows considering some uncertainty in patient clustering compared to the longitudinal-clustering approaches where a patient belongs to a single longitudinal-cluster.

The modified silhouette score also showed comparable performances between the two cluster-tracking approaches. However, it is to note that the network-based cluster-tracking approach does not require the number of clusters to be defined *a priori*. This is an advantage as the number of clusters might be a parameter difficult to set-up. In addition, the network-based cluster-tracking approach has also the advantage of preserving privacy because the interactions between patients are considered rather than absolute data. Another advantage is the flexibility of this approach, as many different measures can be used to compute the similarity between patients. These similarity measures can then be tuned depending on the data and question at hand. Moreover, a large number of algorithms exist for clustering networks.

Code Availability

The code for our two cluster-tracking approaches is available on GitHub (`https://github.com/JudithLamb/Cluster-tracking`). For privacy reasons, antithrombotic drug reimbursements extracted from the EGB cannot be shared publicly. We hence generated a simulated dataset of 5594 patients with their drug use from these extracted data. The results obtained from this simulated sample dataset can be visualized in an R Shiny app also available from the GitHub repository.

Funding

This work was supported by the Inserm cross-cutting program Genomic variability 2018 GOLD.

Acknowledgments

The authors would like to acknowledge Pierre Sabatier for extracting and formatting the data. The authors would also like to thank Anthony Baptista for his contribution in the Methods section. We would like to thank David Hirst, Céline Chevalier, Morgane Terezol and Ozan Ozisik for their many

comments after proofreading the article. And finally, we would like to thank all the members of MMG and Heka teams for their feedback.

References

[1] Cristina Mazzali and Piergiorgio Duca. "Use of administrative data in healthcare research". In: *Internal and emergency medicine* 10.4 (2015), pp. 517–524.

[2] Ivo D Dinov. "Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data". In: *Gigascience* 5.1 (2016), s13742–016.

[3] Sula Windgassen et al. "The importance of cluster analysis for enhancing clinical practice: an example from irritable bowel syndrome". In: *Journal of Mental Health* 27.2 (2018), pp. 94–96.

[4] Anna Okula Basile and Marylyn DeRiggi Ritchie. "Informatics and machine learning to define the phenotype". In: *Expert review of molecular diagnostics* 18.3 (2018), pp. 219–226.

[5] T Warren Liao. "Clustering of time series data—a survey". In: *Pattern recognition* 38.11 (2005), pp. 1857–1874.

[6] Jean-Baptiste Pingault et al. "Childhood trajectories of inattention and hyperactivity and prediction of educational attainment in early adulthood: a 16-year longitudinal population-based study". In: *American Journal of Psychiatry* 168.11 (2011), pp. 1164–1170.

[7] Adeline Divoux et al. "Fibrosis in human adipose tissue: composition, distribution, and link with lipid metabolism and fat mass loss". In: *Diabetes* 59.11 (2010), pp. 2817–2825.

[8] Xiaozhe Wang, Kate Smith, and Rob Hyndman. "Characteristic-based clustering for time series data". In: *Data mining and knowledge Discovery* 13.3 (2006), pp. 335–364.

[9] Daniel S Nagin and Candice L Odgers. "Group-based trajectory modeling in clinical research". In: *Annual review of clinical psychology* 6 (2010), pp. 109–138.

[10] Moritz Herle et al. "Identifying typical trajectories in longitudinal data: modelling strategies and interpretations". In: *European journal of epidemiology* 35.3 (2020), pp. 205–222.

[11] Pablo A Mora et al. "Distinct trajectories of perinatal depressive symptomatology: evidence from growth mixture modeling". In: *American journal of epidemiology* 169.1 (2009), pp. 24–32.

[12] Craig R Colder et al. "Identifying trajectories of adolescent smoking: an application of latent growth mixture modeling." In: *Health Psychology* 20.2 (2001), p. 127.

[13] Aron S Downie et al. "Trajectories of acute low back pain: a latent class growth analysis". In: *Pain* 157.1 (2016), pp. 225–234.

[14] Rebecca J Landa et al. "Latent class analysis of early developmental trajectory in baby siblings of children with autism". In: *Journal of Child Psychology and Psychiatry* 53.9 (2012), pp. 986–996.

[15] Lucas Vendramin, Ricardo JGB Campello, and Eduardo R Hruschka. "Relative clustering validity criteria: A comparative overview". In: *Statistical analysis and data mining: the ASA data science journal* 3.4 (2010), pp. 209–235.

[16] Steven J Van Laere et al. "Uncovering the molecular secrets of inflammatory breast cancer biology: an integrated analysis of three distinct affymetrix gene expression datasets". In: *Clinical cancer research* 19.17 (2013), pp. 4685–4696.

[17] Lovisa Lovmar et al. "Silhouette scores for assessment of SNP genotype clusters". In: *BMC genomics* 6.1 (2005), pp. 1–6.

[18] Victor M Vergara et al. "Determining the number of states in dynamic functional connectivity using cluster validity indexes". In: *Journal of neuroscience methods* 337 (2020), p. 108651.

[19] Jordi A Matias-Guiu et al. "Clustering analysis of FDG-PET imaging in primary progressive aphasia". In: *Frontiers in aging neuroscience* 10 (2018), p. 230.

[20] Yanchi Liu et al. "Understanding and enhancement of internal clustering validation measures". In: *IEEE transactions on cybernetics* 43.3 (2013), pp. 982–994.

[21] Zuyun Liu et al. "Joint trajectories of cognition and frailty and associated burden of patient-reported outcomes". In: *Journal of the American Medical Directors Association* 19.4 (2018), pp. 304–309.

[22] Tracy Vaillancourt and John D Haltigan. "Joint trajectories of depression and perfectionism across adolescence and childhood risk factors". In: *Development and psychopathology* 30.2 (2018), pp. 461–477.

[23] Mitzi M Gonzales et al. "Joint trajectories of cognition and gait speed in Mexican American and European American older adults: The San Antonio longitudinal study of aging". In: *International journal of geriatric psychiatry* 35.8 (2020), pp. 897–906.

[24] William Fung et al. "Joint trajectories of disease activity, and physical and mental health-related quality of life in an inception lupus cohort". In: *Rheumatology* 59.10 (2020), pp. 3032–3041.

[25] Narimene Dakiche et al. "Tracking community evolution in social networks: A survey". In: *Information Processing & Management* 56.3 (2019), pp. 1084–1102.

[26] Derek Greene, Donal Doyle, and Padraig Cunningham. "Tracking the evolution of communities in dynamic social networks". In: *2010 international conference on advances in social networks analysis and mining*. IEEE. 2010, pp. 176–183.

[27] Yang Sun et al. "Matrix based community evolution events detection in online social networks". In: *2015 IEEE international conference on smart city/SocialCom/SustainCom (SmartCity)*. IEEE. 2015, pp. 465–470.

[28] Li Li et al. "Identification of type 2 diabetes subgroups through topological analysis of patient similarity". In: *Science translational medicine* 7.311 (2015), 311ra174–311ra174.

[29] Bo Wang et al. "Similarity network fusion for aggregating data types on a genomic scale". In: *Nature methods* 11.3 (2014), p. 333.

[30] Shraddha Pai and Gary D Bader. "Patient similarity networks for precision medicine". In: *Journal of molecular biology* 430.18 (2018), pp. 2924–2938.

[31] Sarvenaz Choobdar et al. "Assessment of network module identification across complex diseases". In: *Nature methods* 16.9 (2019), pp. 843–852.

[32] Santo Fortunato. "Community detection in graphs". In: *Physics reports* 486.3-5 (2010), pp. 75–174.

[33] Stijn vanDongen. "A cluster algorithm for graphs". In: *Information Systems [INS]* R 0010 (2000).

[34] J MacQueen. "Classification and analysis of multivariate observations". In: *5th Berkeley Symp. Math. Statist. Probability*. 1967, pp. 281–297.

[35] Peter J Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of computational and applied mathematics* (1987), pp. 53–65.

[36] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. "Time-series clustering–a decade review". In: *Information Systems* 53 (2015), pp. 16–38.

[37] Christophe Genolini et al. "kml and kml3d: R packages to cluster longitudinal data". In: *Journal of Statistical Software* 65.4 (2015), pp. 1–34.

[38] Christophe Genolini, Hélène Jacqmin-Gadda, et al. "Copy mean: a new method to impute intermittent missing values in longitudinal studies". In: *Open Journal of Statistics* 3.04 (2013), p. 26.

[39] Alex Nanopoulos, Rob Alcock, and Yannis Manolopoulos. "Feature-based classification of time-series data". In: *International Journal of Computer Research* 10.3 (2001), pp. 49–61.

[40] Tadeusz Caliński and Jerzy Harabasz. "A dendrite method for cluster analysis". In: *Communications in Statistics-theory and Methods* 3.1 (1974), pp. 1–27.

[41] Krzysztof Kryszczuk and Paul Hurley. "Estimation of the number of clusters using multiple clustering validity indices". In: *International workshop on multiple classifier systems*. Springer. 2010, pp. 114–123.

[42] Siddheswar Ray and Rose H Turi. "Determination of number of clusters in k-means clustering and application in colour image segmentation". In: *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*. Citeseer. 1999, pp. 137–143.

[43] David L Davies and Donald W Bouldin. "A cluster separation measure". In: *IEEE transactions on pattern analysis and machine intelligence* 2 (1979), pp. 224–227.

[44] Hirotogu Akaike. "Information theory and an extension of the maximum likelihood principle". In: *Selected papers of hirotugu akaike*. Springer, 1998, pp. 199–213.

[45] Gideon Schwarz. "Estimating the dimension of a model". In: *The annals of statistics* (1978), pp. 461–464.

[46] Ron Wehrens, Hein Putter, and Lutgarde MC Buydens. "The bootstrap: a tutorial". In: *Chemometrics and intelligent laboratory systems* 54.1 (2000), pp. 35–52.

[47] P1 Tuppin et al. "French national health insurance information system and the permanent beneficiaries sample". In: *Revue d'epidemiologie et de sante publique* 58.4 (2010), pp. 286–290.

[48] Armin Skrbo, Begler Begović, and Selma Skrbo. "Classification of drugs using the ATC system (Anatomic, Therapeutic, Chemical Classification) and the latest changes". In: *Medicinski arhiv* 58.1 Suppl 2 (2004), pp. 138–141.

[49] Jin Liu and Robert A Perera. "Extending Growth Mixture Model to Assess Heterogeneity in Joint Development with Piecewise Linear Trajectories in the Framework of Individual Measurement Occasions". In: *arXiv preprint arXiv:2010.13325* (2020).

[50] Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*. John Wiley & Sons, 1991.

[51] Juhani Knuuti et al. "2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromes: The Task Force for the diagnosis and management of chronic coronary syndromes of the European Society of Cardiology (ESC)". In: *European heart journal* 41.3 (2020), pp. 407–477.

[52] Victor Aboyans et al. “ESC Scientific Document Group. 2017 ESC Guidelines on the Diagnosis and Treatment of Peripheral Arterial Diseases, in collaboration with the European Society for Vascular Surgery (ESVS): Document covering atherosclerotic disease of extracranial carotid and vertebral, mesenteric, renal, upper and lower extremity arteriesEndorsed by: the European Stroke Organization (ESO) The Task Force for the Diagnosis and Treatment of Peripheral Arterial Diseases of the European Society of Cardiology (ESC) and of the European Society for Vascular Surgery (ESVS)”. In: *Eur Heart J* 39.9 (2018), pp. 763–816.

[53] Marco Valgimigli et al. “2017 ESC focused update on dual antiplatelet therapy in coronary artery disease developed in collaboration with EACTS: The Task Force for dual antiplatelet therapy in coronary artery disease of the European Society of Cardiology (ESC) and of the European Association for Cardio-Thoracic Surgery (EACTS)”. In: *European heart journal* 39.3 (2018), pp. 213–260.

[54] Robert G Hart, Lesly A Pearce, and Maria I Aguilar. “Meta-analysis: antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation”. In: *Annals of internal medicine* 146.12 (2007), pp. 857–867.

[55] Gerhard Hindricks et al. “2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS) The Task Force for the diagnosis and management of atrial fibrillation of the European Society of Cardiology (ESC) Developed with the special contribution of the European Heart Rhythm Association (EHRA) of the ESC”. In: *European heart journal* 42.5 (2021), pp. 373–498.

[56] Deirdre P Cronin-Fenton et al. “Hospitalisation for venous thromboembolism in cancer patients and the general population: a population-based cohort study in Denmark, 1997–2006”. In: *British journal of cancer* 103.7 (2010), pp. 947–953.

[57] Emmanuel Oger, EPI-GETBO study group, et al. “Incidence of venous thromboembolism: a community-based study in Western France”. In: *Thrombosis and haemostasis* 83.05 (2000), pp. 657–660.

[58] Karine Chevreul, Karen Berg Brigham, and Clara Bouché. “The burden and treatment of diabetes in France”. In: *Globalization and health* 10.1 (2014), pp. 1–9.