THE SCOURGE OF ANTIMICROBIAL RESISTANCE: A MACHINE LEARNING APPROACH FOR PRESCRIPTION PATTERNS ANALYTICS

Davide Castaldi(1), Ilaria Giordani(2), Antonio Candelieri(2), Roberto Mattina(3)
Francesco Archetti(1),(2)

(1) Consorzio Milano Ricerche
Via Roberto Cozzi 53, 20125, Milan, Italy
castaldi@milanoricerche.it
archetti@milanoricerche.it

(2) University of Milano-Bicocca, Department of Computer Science, Systems and Communication
Viale Sarca 336, U14 Building, 20126, Milan, Italy
ilaria.giordani@disco.unimib.it
antonio.candelieri@unimib.it
francesco.archetti@unimib.it

(3) University of Milano, Department of Biomedical, Surgical and Dental Sciences
Via Pascal, 36/38, 20133, Milan, Italy
roberto.mattina@unimi.it

Keywords: timeseries clustering, data analytics, prescription patterns, antimicrobial resistance

Abstract. Antimicrobial resistance (AMR) has become a health emergency worldwide: bacteria are mutating and exchanging their genes at an increasing rate, with 5% growth for some bacteria resistant to some classes of antibiotics in some EU countries [1]. Data analytics can give a major contribution to tackle this health and socioeconomic challenge: by means of insights originated from analytical results health authorities can plan informed actions.

This study summarizes relevant insights obtained by the analysis of data related to about 500 general practitioners (GPs) and around one million patients, during the period 2011-2016. After specific data cleaning, the sample was reduced to just over 500’000 patients, balanced in terms of gender, age, and therapeutic indication, and related to 140 different types of anti-bacterial agents.

We employed a time series clustering approach to first identify the typical prescription patterns and then discriminating between GPs with changing and constant prescription behaviors. More precisely, spherical k-means clustering, based on cosine similarity, was used to obtain clusters consisting of prescription patterns sharing occurrence of peaks and bursts. Finally, it was also possible to identify possible changes in prescription patterns, from one year to the next, for every GP as well as for similar groups of GPs. The approach proposed in this paper provides useful insights to policy makers and health care actors about local trend or

individual behaviors enabling the design and implementation of consistent and targeted corrective interventions. The overuse, underuse or misuse of antibiotics molecules results in wastage of relevant resources to challenge infections, contributing to health hazards spread related to AMR.

1 Scientific Background

The history of drug resistance began with the development of antimicrobial drugs, and the subsequent ability of microbes to adapt to survive in the presence of antimicrobials. Increasing global consumption of antibiotics, their inappropriate prescription, as well as international travelling rising, push bacteria to change and interchange their genetic heritage at an extraordinary rate. Over the last twenty years, antibiotic selective pressure and facilitated colonization conditions are main features contributing to the dissemination of antibiotic resistance worldwide. Although most antibiotic-resistant bacteria originally emerged in hospital settings, drug-resistant strains are becoming increasingly common in the community.

Moreover, multi-resistant bacteria are a worldwide critical public health issue that requires a coordinated and multifaceted response, not excluding veterinarians and farmers, as reconfirmed in 2018 by World Health Organization (WHO) [2].

Currently, more than in other developed Countries, AMR is a real issue for Italian public health, especially for Southern Regions [3], that require focused campaigns able to catch attention of all health care actors, but also able to reach and sensitize the public.

According to insights originating from analytical approaches, relatively to prescription behaviors, key players can plan, schedule and intervene with informed decisions on territory, both by therapeutic recommendations (suggesting more effective molecules, therapeutic switch or alternatives) and GP education campaign on prescriptions appropriateness. Some relevant works about the analysis of time series of medical prescriptions are reported in [4] and [5]. The former proposes to train an artificial neural network on Electronic Health Record (EHR) data with the aim to support decisions about prescriptions. The latter proposes an interrupted time series regression analysis to evaluate modifications of quinolone antibiotics prescriptions induced by the introduction of a novel restricted reimbursement policy. The studies considered patient-centered data, while the contribution of this paper is to analyze prescriptions explicitly associated to GPs, allowing to evaluate the individual prescription behavior, even over the time.

2 Materials and Methods

2.1 The database at a glance

The available database was collected, managed and complied with current privacy regulations, by the “Consorzio Nazionale delle Cooperative Mediche” (CNCM) before being delivered for our analysis. It is composed of pseudonymised data by encryption before our acquisition and is related to seventeen years data about outpatient visits at GPs’ ambulatories located in several Health Districts within the Campania region in the southern Italy. The overall number of observed patients is 1.042.321 (updated to June 2017), representing just under 20% of residents of this region. In more than 17 years, from 2000 to June 2017, were collected: 80.000.000 visits, 115.522.740 therapy prescriptions and 137.463.306 clinical ascertainment requests.

Registry patient's data reveal 472.739 (47.20%) males, 527.363 (52.66%) females and 1378 (0.14%) unspecified. These gender data are comparable to the data reported by the Italian National Institute of Statistics for the year 2016 (http://www.istat.it/en/), confirming the representativeness of the database.

Age distribution shows a 4.30% (43.085) of pediatric patients, ranging from 0 to 17 years, 66.30% adult (663.942) and 29.11% (291.561) elderly people, and nearly 0.3% without birth date filled out.

2.2 Antibiotic Therapy dataset

To analyze antibiotic therapy prescription patterns, a goal specific dataset was extracted. The observation period was limited to the window 2011-01-01 to 2016-12-31. (2017 has been excluded because of incompleteness, as the last update was in June 2017). This selection is in accordance with international health reports about this public health emergency [6]. The resulting dataset consists of 550.232 patients having at least one therapeutic indication for antibiotics in the selected period (54.94% of the patients in the original database). With respect to gender, 252.244 patients are male (45.84%) and 297.700 are female (54.11%), while gender information is missing for 288 patients (0.05%). With respect to age, and depending on gender, the dataset consists of:

- Male patients: 12.470 (4.94%) pediatric, 170.889 (67.75%) adults and 68.866 (27.30%) elderly.
- Female patients: 14.228 (4.78%) pediatric, 193.504 (65.00%) adults and 89.945 (30.22%) elderly

Antibiotics prescribed during the six years analyzed (2011-2016) include 140 different types of antibacterial agents identified by the Anatomical Therapeutic Chemical code (ATC code) corresponding to 4.111.556, from which, most prescribed antibiotics belong to ATC class “J01” with 3.688.304 (89.70%) prescriptions and “A07AA” with 409.057 (9.95%) prescriptions, respectively antibacterial for systemic use and intestinal infections.

2.3 Machine Learning Approach

Clustering algorithms belong to the family of *unsupervised* machine learning approaches. The final goal of any clustering algorithm is to group data by maximizing the similarity of samples assigned to the same group/cluster (i.e., maximizing intra-cluster similarity) while minimizing the similarity between samples assigned to different groups/clusters (i.e., minimizing inter-cluster similarity). Most of the clustering applications deal with *not-sequential data*, where every sample is a data point in the *n*-dimensional space spanned by the features (also known as “attributes”) describing the samples of the available data set. Although the aim of clustering remains the same when time series data is considered, the order of the features has its own meaning, representing the sequence of observations over time. This requires specific choices for data representation, pre-processing, and selection of the similarity measure to use [7].

With respect to data representation, and according to the goal of the analysis proposed in this paper, we chose to work with the "raw" representation of the data, where every sample is represented as a 12-dimensional vector. More precisely, every component of these vectors is the number of monthly prescriptions. Every sample refers to a specific year in the observation period (2011-2016) and a GP.

A useful characterization of similarity measures, used to compare and cluster time-series data, was proposed in [8]:

- Similarity in time. Comparison of time series considers how two time series vary at each time step. In this case, time series can be clustered by capturing repetitive behaviours occurring always at the same time step or in the same time window (e.g., peaks/bursts of monthly prescriptions)
- Similarity in shape. Comparison of time series considers common shape features, such as trends, even occurring at different time step, as well as similar sub-patterns.
- Similarity in change. Comparison of time series considers variation from time step to time step, allowing for grouping sample with same variations between two successive time stamps.

For the aim of this study, that is identifying typical prescription behaviours, we adopted a "similarity in time" measure, more precisely the cosine similarity. Cosine similarity is given by the cosine of the angle between two vectors, so it ranges in the interval $[-1;1]$. Following, the formula of cosine similarity is reported, where $\langle x_i, x_j \rangle$ represents the scalar product between two vectors (i.e., time series), x_i and x_j, while $\| \quad \|$ is the module of the vector.

$$s(x_i, x_j) = \frac{\langle x_i, x_j \rangle}{\|x_i\| \, \|x_j\|}$$

As the components of the prescriptions vectors are not negative, the possible values of the cosine similarity is restricted to $[0;1]$.

For clustering, we used the spherical k-means algorithm provided by the R package "skmeans", which implements a simple k-means strategy using a distance measure based on the cosine similarity:

$$(x_i, x_j) = 1 - s(x_i, x_j) = 1 - \frac{\langle x_i, x_j \rangle}{\|x_i\| \, \|x_j\|}$$

A similar approach was already applied to identify typical water consumption patterns in urban water distribution networks [9][10].

As any k-means clustering algorithms, also spherical k-means requires to set up in advance the number k of desired clusters. However, it is usually difficult to identify the best value of k a-priori. In this study, we considered k varying from 3 to 20. Moreover, we performed 30 different runs of the spherical k-means, for each value of k, with the aim to mitigate the effect of randomness of the clustering initialization. Than we selected most suitable value for k, according to the average value of the Silhouette index. The Silhouette index measures of how similar a sample is to the others belonging to the same cluster (cohesion) compared to how similar it is to the other samples in the other clusters (separation). Silhouette ranges from −1 to +1, where a high value indicates that the sample is well matched to its own cluster and poorly matched to the others. The result of a clustering algorithm is appropriate when samples have high Silhouette values.

3 Results

As a first result, the boxplot of the Silhouette index computed for every value of *k* is reported in Figure 1. According to the boxplot, the Silhouette index increases, on average, with *k*, but only up to *k*=12. When *k* becomes greater than 12 the Silhouette index decreases drastically. This means that 12 is the most suitable number of clusters to group the prescriptions time series in the dataset.

Figure 2 reports the 12 time series that are the representative (cluster prototype) for each cluster (i.e., average time series computed on each cluster). Since the cosine similarity is a "similarity in time", it considers similar those time series having peaks/bursts at the same time step. This is clearly visible in Figure 2, where every time series in the picture has a peak in prescriptions at specific month of the year depending on the cluster.

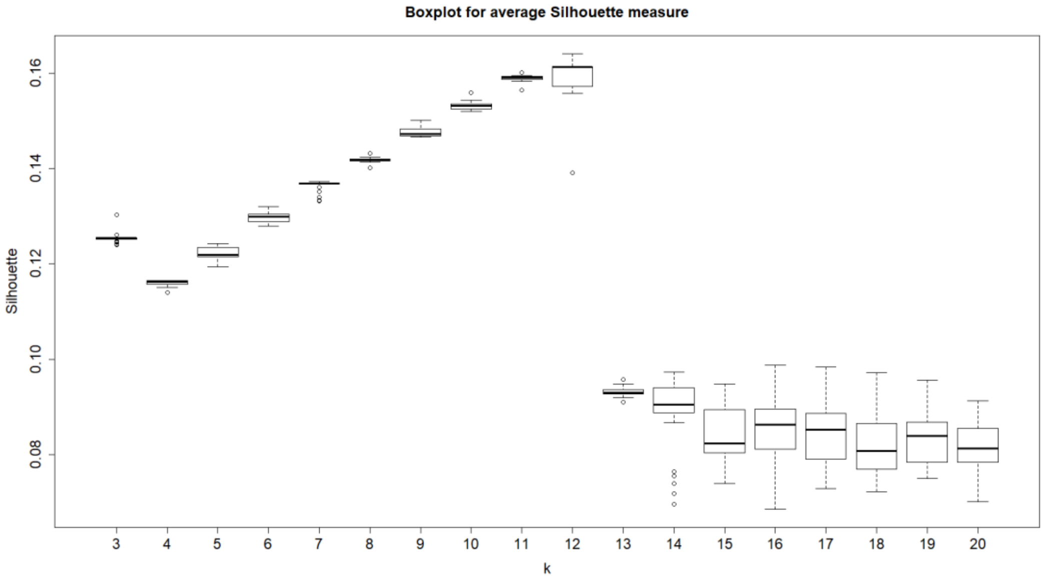


Figure 1. Boxplot of Silhouette index for different values of *k* in the spherical *k*-means algorithm. Boxes are computed from 30 different runs of the algorithm for every k value

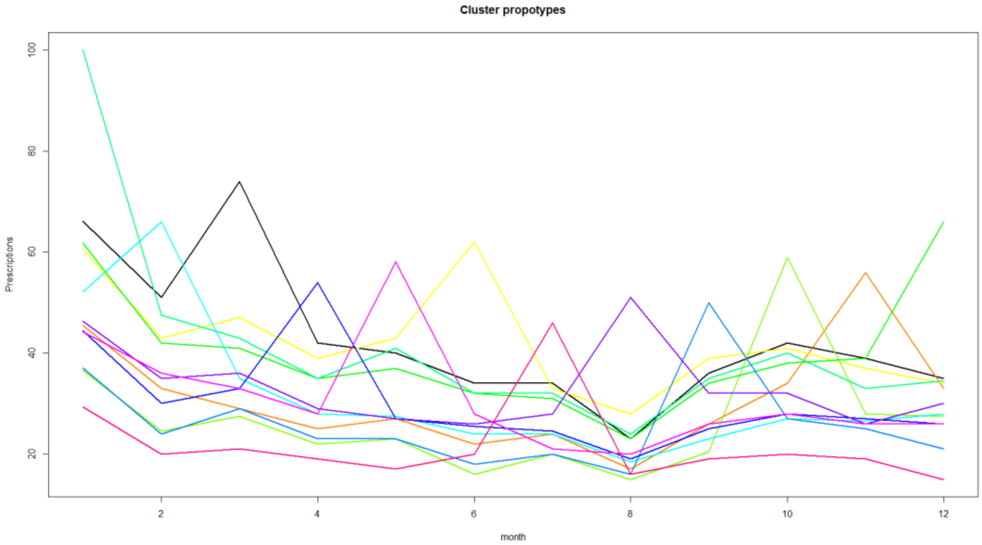


Figure 2. The 12 time series representing the 12 clusters. Every time series is computed as the average of all the time series belonging to that cluster (cluster prototype) and summarizes a prescription behavior

The following Table 1 summarizes the size of clusters (i.e., number of time series belonging to each cluster). As result, sizes are substantially balanced; this permits to conclude that the 12 prescription behaviors identified are quite “typical” and that there are not significantly “anomalous” behaviors.

Table 1. Size of every cluster

Cluster	1	2	3	4	5	6	7	8	9	10	11	12
Size	417	231	273	182	254	384	290	201	230	171	251	115

The following Figure 3 reports an example about the time series composing two different clusters, respectively cluster 4 and cluster 6, and their prototypes. Grey lines are the time series belonging to the cluster while the black line is the cluster’s prototype. Cluster 6 (on the right) has a higher variance, even due to the higher number of time series associated (384 vs 182), but the most relevant difference between the two clusters is the peak in the prescriptions patterns, in October and in January, for cluster 4 and 6, respectively.

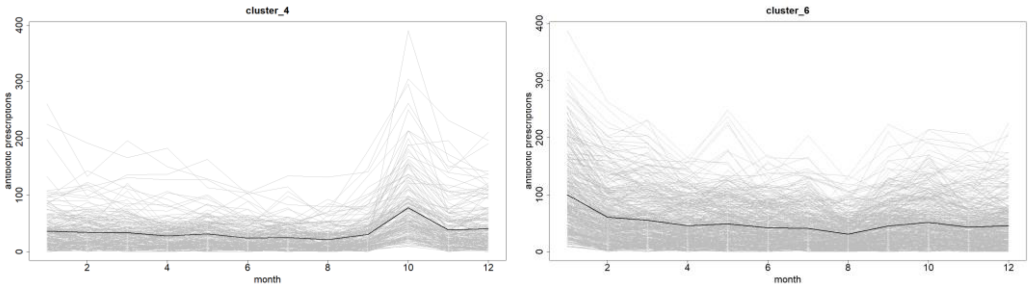
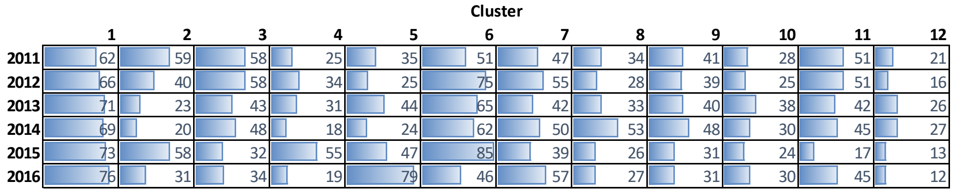


Figure 3. Composition of two clusters: grey lines are antibiotic prescription time series, with monthly step, assigned to the cluster, black line is the cluster’s prototype

Although prescription patterns, within a cluster, can be centered on different offsets, it is important to remind that spherical k-means identifies clusters just according to similarity in time, by considering the angle between vectors, and therefore results are quite like those obtained by performing a preliminary data normalization.
After the creation of the clusters we performed a check on possible relations between prescription behaviors and years of the observation period. The following Figure 4 reports the number of time series in each couple “year-cluster”. Some relevant insights can be inferred: *cluster_1* (and the associated prescription pattern) slowly grow over time, while *cluster_3* was quickly disappearing. Moreover, *cluster_2* was unusually common in 2011 and 2015, *cluster_8* in 2014, *cluster_4* in 2015, *cluster_5* in 2016, while *cluster_11* was greatly abandoned in 2015.
Although it would be very interesting to link the occurrence of these patterns, in the observation period and area, to significant events (e.g. microbial or disease surges), the relevant data were not available for this study.



	Cluster											
	1	2	3	4	5	6	7	8	9	10	11	12
2011	62	59	58	25	35	51	47	34	41	28	51	21
2012	66	40	58	34	25	75	55	28	39	25	51	16
2013	71	23	43	31	44	65	42	33	40	38	42	26
2014	69	20	48	18	24	62	50	53	48	30	45	27
2015	73	58	32	55	47	85	39	26	31	24	17	13
2016	76	31	34	19	79	46	57	27	31	30	45	12

Figure 4. Distribution of time series data with respect to clusters and years

According to our clustering, 10% of GPs in the dataset have not modified their prescription behavior over the observation period. Moreover, these GPs have not all the same behavior: the following Figure summarizes the distribution of the “GPs with constant prescription behavior” with respect to the clusters identified.

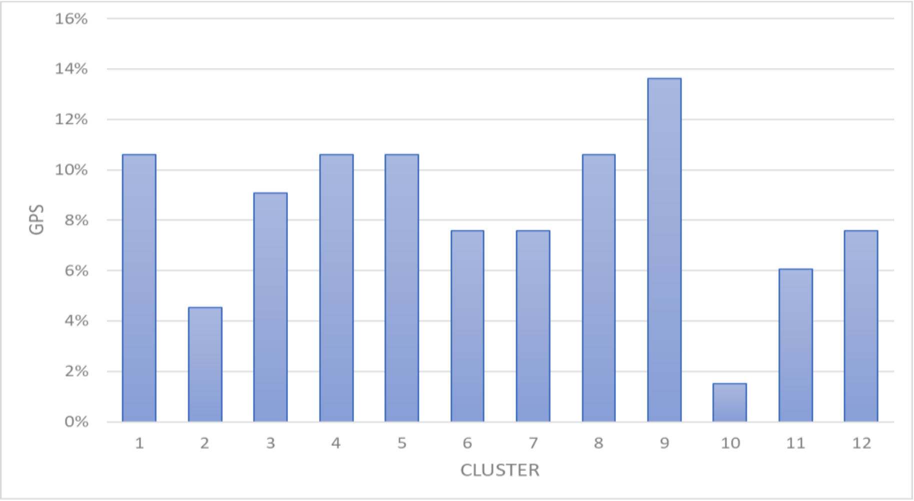


Figure 4. Percentage distribution of the GPs who did not change prescription behavior over time, with respect to the clusters identified

Finally, merging all the afore mentioned insights, is possible to assess: how many typical prescription behaviors can be observed over a period – and area – of interest, how their frequency changes over time – and possibly correlate that with relevant events – and finally identify an association between GPs and typical prescription behaviors and, even more important, changes in individual prescription behavior over time (e.g., compliantly with events and or introduction of new policies).

4 Conclusions

A key objective of this research was to identify typical patterns in the antibiotic prescription behavior using a relevant set of Italian GPs prescription data. The merge of the different insights resulting from the proposed approach can effectively support the assessments of National or Regional social-health policies and, specifically at local level, the evaluation of prescription behaviors and their modifications. Thus, it can enable mechanisms for monitoring compliance to recommendations of Territorial Social-Health Institutions and Health Protection Agency (respectively Italian ASST and ATS) aimed at reducing inappropriate prescribing, dispensing and use of antimicrobials, which is one of main factors inducing AMR. Finally, a critical evaluation of the time series clustering results can provide a set of valuable insights about antibiotic resistance surveillance, in accordance with concept of "One health", which is the recommended perspective for antimicrobial resistance management [2].

References

[1] ECDC. Surveillance of antimicrobial resistance in Europe. Annual Report of the European Antimicrobial Resistance Surveillance Network (EARS-Net). (2017).

[2] Sabiha, E. Antibiotic resistance and One Health: a mapping project. The Lancet Global Health, 6(S27). (2018).

[3] Prigitano, A., Romanò, L., Auxilia, F., Castaldi, S. & Tortorano, A.M. Antibiotic resistance: Italian awareness survey 2016. Journal of Infection and Public Health. 11(1), 30-34. (2017).

[4] Xia, E., Mei, J., Xie, G., Li, X., Li, Z. & Xu, M. Learning Doctors' Medicine Prescription Pattern for Chronic Disease Treatment by Mining Electronic Health Records: A Multi-Task Learning Approach. AMIA Annu Symp Proc, 1828–1837. (2018).

[5] Manns, B., Laupland, K., Tonelli, M., Gao, S. & Hemmelgarn, B. Evaluating the impact of a novel restricted reimbursement policy for quinolone antibiotics: a time series analysis. BMC Health Serv Res. 12:290. (2012).

[6] Klein, E.Y., Van Boeckel, T.P., Martinez, E.M., Pant, S., Gandra, S., Levin, S.A., Goossens, H. & Laxminarayan, R. Global increase and geographic convergence in antibiotic consumption between 2000 and 2015. Proc Natl Acad Sci USA. 115(15), E3463-E3470. (2015).

[7] Kavitha, V. & Punithavalli, M. Clustering time series data stream-a literature survey. Int J Comput Sci Inf Secur IJCSIS 8(1), 289-294. (2010).

[8] Zhang, X., Liu, J., Du, Y. & Lv, T.: A novel clustering method on time series data. Expert Syst. Appl., 38, 11891–11900. (2011).

[9] Shabani, S., Candelieri, A., Archetti, F. & Naser, G. Gene Expression Programming Coupled with Unsupervised Learning: A Two-Stage Learning Process in Multi-Scale, Short-Term Water Demand Forecasts. Water, 10(2), 142. (2018).

[10] Candelieri, A. Clustering and support vector regression for water demand forecasting and anomaly detection. Water, 9(3), 224. (2017).