IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received July 6, 2020, accepted July 15, 2020, date of publication July 27, 2020, date of current version August 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3012082

Data-Driven Visual Characterization of Patient Health-Status Using Electronic Health Records and Self-Organizing Maps

DAVID CHUSHIG-MUZO[1], **CRISTINA SOGUERO-RUIZ**[1], **A. P. ENGELBRECHT**[2,3], **(Senior Member, IEEE), PABLO DE MIGUEL BOHOYO**[4], **AND INMACULADA MORA-JIMÉNEZ**[1]

[1]Department of Signal Theory and Communications, Telematics and Computing Systems, Rey Juan Carlos University, 28943 Fuenlabrada, Spain
[2]Department of Industrial Engineering, Stellenbosch University, Stellenbosch 7600, South Africa
[3]Computer Science Division, Stellenbosch University, Stellenbosch 7600, South Africa
[4]University Hospital of Fuenlabrada, 28943 Fuenlabrada, Spain

Corresponding author: Inmaculada Mora-Jiménez (inmaculada.mora@urjc.es)

This work was partly supported by the Spanish Research projects PID2019-106623RB-C41 (Beyond), TEC2016-75361-R (Klinilycs), PID2019-107768RA-I00 (AAVis-BMR), DTS17/00158, and TIN2017-90567-REDT, by the Eramus Mundus Program (Action 2 Partnerships, AESOP+, Agreement Number 2014-0878 / 001-001), by Project Ref. F656 funded by Rey Juan Carlos University and by the Young Researchers R&D Project Ref. 2020-661 financed by Rey Juan Carlos University and Community of Madrid (Spain).

ABSTRACT Hypertension and diabetes have become a global health and economic issue, being among the major chronic conditions worldwide, particularly in developed countries. To face this global problem, a better knowledge about these diseases becomes crucial to characterize chronic patients. Our aim is two-fold: (1) to provide an efficient visual tool for identifying clinical patterns in high-dimensional data; and (2) to characterize the patient health-status through a data-driven approach using electronic health records of healthy, hypertensive and diabetic populations. We propose a two-stage methodology that uses diagnosis and drug codes of healthy and chronic patients associated to the University Hospital of Fuenlabrada in Spain. The first stage applies the Self-Organizing Map on the aforementioned data to get a set of prototype patients which are projected onto a grid of nodes. Each node has associated a prototype patient that captures relationships among clinical characteristics. In the second stage, clustering methods are applied on the prototype patients to find groups of patients with a similar health-status. Clusters with distinctive patterns linked to chronic conditions were found, being the most remarkable highlights: a cluster of pregnant women emerged among the hypertensive population, and two clusters of diabetic individuals with significant differences in drug-therapy (insulin and non-insulin dependant). The proposed methodology showed to be effective to explore relationships within clinical data and to find patterns related to diabetes and hypertension in a visual way. Our methodology raises as a suitable alternative for building appropriate clinical groups, becoming a promising approach to be applied to any population due to its data-driven philosophy. A thorough analysis of these groups could spawn new and fruitful findings.

INDEX TERMS Electronic health records, machine learning, self organizing maps, clustering, data visualization, chronic conditions.

I. INTRODUCTION

In recent years, the number of chronic patients is increasing at an alarming rate, becoming a public health concern at a global scale. Treating patients with chronic diseases generally increases the need to face a growing healthcare demand. Several reports released by the World Health Organization [1], [2] (WHO) highlight the importance about prevention and changes in policies to reduce health risks associated to chronic conditions. Characterization of patient health-status and its evolution become a priority to prevent the onset of chronic diseases, to improve the patient's quality

The associate editor coordinating the review of this manuscript and approving it for publication was Juan A. Lara.

of life, to reduce costs, and to make efficient use of healthcare resources [3], [4].

Data-driven models based on Machine Learning (ML) have been intensively considered for extracting knowledge and discovering patterns related to diseases in different clinical works [5]–[9]. ML is a scientific discipline which uses learning (experience guided by examples) to build plausible models with a reasonable generalization capability [10].

The widespread adoption of Electronic Health Records (EHRs) has generated unprecedented amounts of digitized data for clinical research, contributing towards the development of many ML-based works [11]–[15]. However, working with clinical data becomes challenging due to the high-dimensionality of the data, since a patient is usually represented by many features. The high-dimensionality makes it difficult to visualize data and to interpret their patterns. To overcome the lack of interpretation, this work uses a ML technique, namely the Self-Organizing Map (SOM) [16]–[18]. The SOM provides a visual way for exploratory data analysis and clustering, having been used in different domains, ranging from pattern recognition, speech recognition, signal processing and finance [19]–[22]. In the clinical domain, the SOM has also been applied to identify patterns in patients with certain diseases, to predict risk of diseases, and to extract knowledge from drug use or identifying groups of patients according to clinical codes, among others [23]–[33].

It is well-known that high-dimensionality in clinical data produces a challenge for medical interpretation, as well as important drawbacks for the extraction of clinical patterns. The SOM has shown maturity to contribute to the data interpretation and to provide novel clinical evidence when applied to health problems [34]–[36]. The SOM maps complex relationships in the data into simple geometric relationships on a low-dimensional space [37], usually a bi-dimensional grid of nodes. The grid structure provides a visual way for exploratory data analysis, paving the way for the discovery of hidden patterns and clustering patients with common clinical characteristics. This is due to its topology-preserving property, which allows projection of similar patients from the high-dimensional space onto the same region of the grid. Each node of the grid has associated a prototype vector (called in this paper *prototype patient*), which is a representation of the clinical characteristics. We used these prototype vectors provided by the SOM as input to the clustering method in order to group prototype patients (thereby patients) with similar clinical characteristics.

This study proposes a two-stage methodology for characterizing patient health-status following a data-driven approach. In the first stage, the SOM is trained, and prototype patient vectors are obtained according to diagnosis and drug codes. In the second stage, clustering methods are performed for grouping 'similar' prototype patients (thereby nodes in the grid) for getting clusters with chronic patterns.

This paper is structured as follows. Section II includes data description and pre-processing used for characterizing the health-status and the patient description through their EHR. Section III describes the theoretical fundamentals of the SOM, as well as the considered clustering methods and cluster validity scores. Experimental results are detailed in Section IV. The discussion and conclusions are presented in Section V and Section VI, respectively.

II. DATA DESCRIPTION AND PRE-PROCESSING

In this section, we describe the dataset as well as the pre-processing stage performed. Further, a visual characterization based on profiles of populations is carried out.

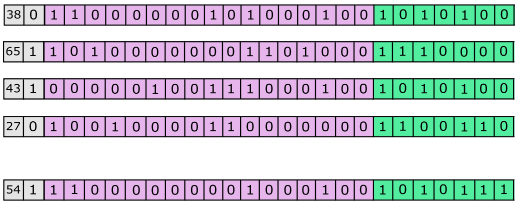
Data have been provided by the University Hospital of Fuenlabrada (UHF) in Madrid, Spain, during one year. The UHF is a public hospital with almost 220,000 citizens assigned, and its activity, per year, is around 420,000 outpatients, 15,500 discharges, 12,000 surgeries and 120,000 emergencies. We work with different features: age, gender, diagnosis codes according to the International Classification of Diseases Ninth Revision-Clinical Modification (ICD9-CM) [38] and pharmaceutical drug codes following the Anatomical Therapeutic Chemical (ATC) Classification System [39]. Both ICD9-CM and ATC codes are recommended by WHO and have been extensively used in a variety of studies at international level [40]–[43].

A. DEMOGRAPHIC CHARACTERISTICS AND CLINICAL CODES

ICD9-CM codes consist of six alphanumeric-characters with a decimal point between the third and fourth character [38]. ATC codes are composed of seven alphanumeric-characters hierarchically structured in five levels: anatomical (first element), therapeutic (second and third element), pharmacological (fourth element), chemical (fifth element) and chemical substance (sixth and seventh element). Following a similar approach in [44], [45] and aiming to decrease the number of features, we reduced the detail of the aforementioned codes. The characters after the decimal point were removed for ICD9-CM and the fifth level of the ATC codes were discarded. These simplified-version codes are named 'short-codes'. This simplification in diagnosis and drug codes results in a patient representation by a vector of 2,265 features (see Fig. 1): age, gender, 1,517 diagnoses and 746 drugs. The features related to diagnoses and drugs are coded by binary values denoting the presence/absence ('1'/'0') of the corresponding diagnosis and drug code for a particular patient during the year of study.

B. CLINICAL RISK GROUPS

ICD9-CM and ATC codes have been used as valuable data for identifying chronic populations in prior studies [46]–[48]. Patient Classification Systems (PCSs) are used as a tool to clinically validate the findings provided by the ML techniques proposed [49]. PCSs apply a set of clinical rules to assign a patient to one group with similar characteristics from a clinical viewpoint. These rules, stated by physicians with broad knowledge and expertise, are extracted by using a high number of clinical records. Among PCSs,



38	0	1	1	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0	1	0	1	0	0
65	1	1	0	1	0	0	0	0	0	0	0	1	1	0	1	0	0	0	1	1	1	0	0	0	0
43	1	0	0	0	0	0	1	0	0	1	1	1	0	0	0	1	0	0	1	0	1	0	1	0	0
27	0	1	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0	1	1	0
54	1	1	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	1	0	1	1	1

FIGURE 1. Representation of patient vectors. Representation of conditions of *N* patients as *N* vectors of features. Each patient is represented by age; gender; 1,517 diagnosis codes (ICD9-CM short-codes) and 746 drug codes (ATC short-codes).

the system named Clinical Risk Groups (CRGs) [50], has been clinically validated at an international level in a variety of works [51]–[53] and is oriented towards the identification of chronic patients. The CRGs use demographic data, diagnoses, clinical procedures and drugs in a period of time to assign each individual into a severity-adjusted homogeneous health-status. Since CRGs provide a categorization clinically accepted for identifying main chronic conditions, it can used to validate the result provided when applying the methodology proposed in this work. Note that this validation can just be used to confirm the clinical knowledge considered by the CRGs rules. Nevertheless, if the main results provided by the ML approach are in accordance with the clinical knowledge, the use of our approach can open the way to new clinical findings. We considered the health-status of the patient through three CRGs: CRG-1000 (healthy) with 46,835 patients, CRG-5192 (hypertensive) with 12,447 patients, and CRG-5424 (diabetic) with 2,166 patients.

In recent works related to chronic conditions [45], [54], we proposed a representation named *profile* to characterize every health-status. For a specific CRG, the profile is an uni-dimensional visual representation where the horizontal axis contains the diagnosis/drug codes and the vertical axis shows the average rate of each code when considering all individuals belonging to that CRG. The profile allows to check the prevalence of certain codes against others, and provides information about which codes are associated with a particular CRG. Fig. 2 depicts the profiles according to each CRG (CRG-1000, CRG-5192 and CRG-5424) taking into account ICD9-CM short-codes (left panels) and ATC short-codes (right panels). For example, looking at the diagnosis profile associated with CRG-5424 (Fig. 2 (e)), the code with the highest value is ICD9-CM '250', which indicates that around 90% of the individuals belonging to this CRG were diagnosed with diabetes.

C. IMBALANCE CLASSES

In general, most ML techniques are affected by imbalance in classes [55]: when the dataset used for learning has a considerable proportion of samples for some classes, the learning process can be monopolized by those associated to the majority group. To solve this issue, several strategies have been proposed in the literature [56]–[58]. For simplicity, we consider the random under-sampling strategy [59], where classes with more samples are under sampled to have the same number of samples than the minority class. Since the classes considered in this work exhibit the imbalance problem (much more individuals with a healthy condition), we balance our dataset taking as reference the minority class, i.e. CRG-5424 with 2,166 patients.

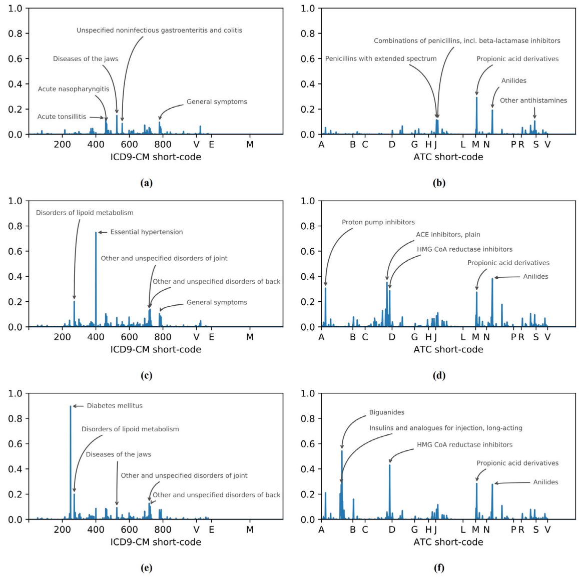


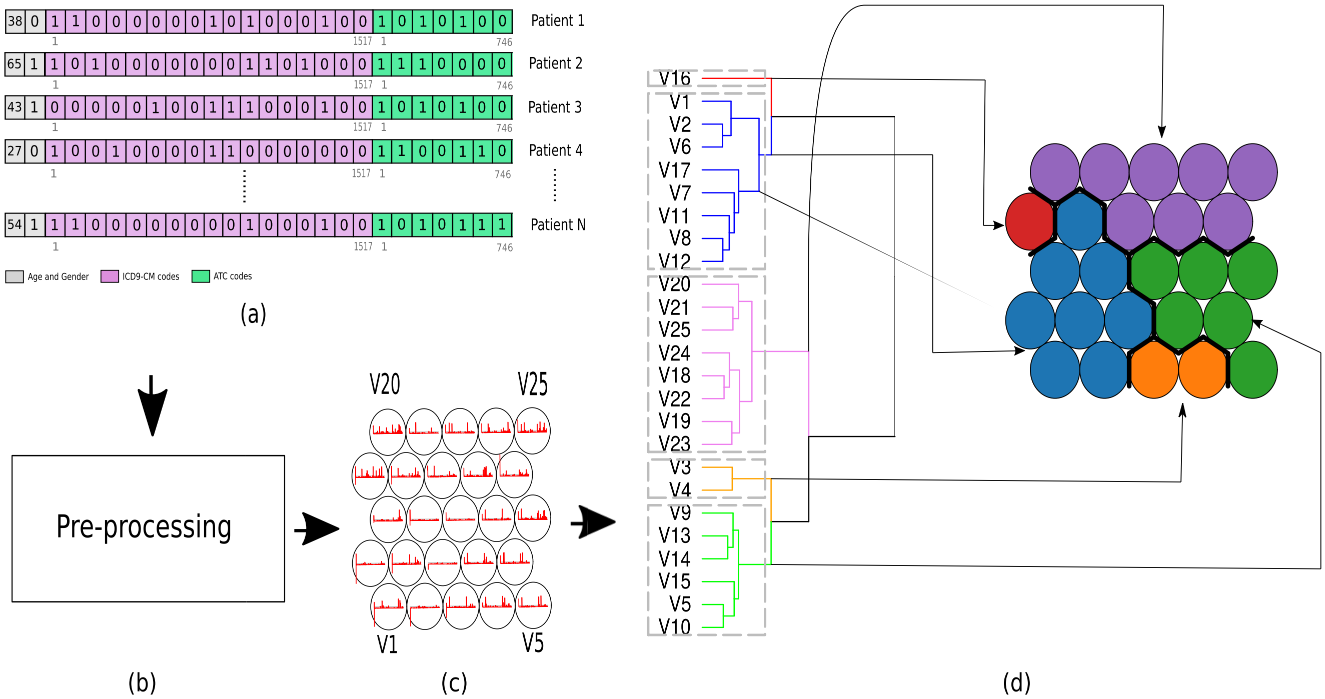
FIGURE 2. Profiles associated to CRGs. Diagnosis profiles (left panels) and drug profiles (right panels) for the CRGs: (a-b) CRG-1000 (healthy patients); (c-d) CRG-5192 (hypertensive patients); (e-f) CRG-5424 (diabetic patients). Note that the five ICD9-CM and ATC short-codes with the highest average rate values are pointed out.

III. SELF-ORGANIZING MAP AND CLUSTERING FOR THE TWO-STAGE METHODOLOGY

This study is carried out by following a two-stage methodology as in [37], [60] (depicted in Fig. 3): the first stage uses SOM to build a set of prototype vectors capturing clinical insights from data, whereas the second stage uses these vectors as input to a clustering method to identify groups of patients with similar characteristics. This section introduces fundamentals of Self-Organizing Map and the clustering methods used in this study.

A. SELF-ORGANIZING MAP

The SOM is a type of artificial neural network based on unsupervised learning proposed by Kohonen [16]–[18], extensively used for exploratory data analysis and visualization of high-dimensional data. The SOM maps the high-dimensional space into a lower-dimensional space formed by a set of M nodes arranged in a grid-structure. Formally, given a set of N samples given by $X = \{x_1, x_2, x_3 \ldots, x_N\}$ (where each vector is characterized by a D-dimensional vector $x_i = [x_{i,1}, \ldots, x_{i,D}]$), the SOM assigns each vector to one node of the grid. Each node j is characterized by a prototype vector $v_j = [v_{j,1}, \ldots, v_{j,D}], j = 1, \ldots, M$ [37]. Nodes of the grid are organized to maintain the topological properties



38	0	1	1	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0	1	0	1	0	0	Patient 1
65	1	1	0	1	0	0	0	0	0	0	0	1	1	0	1	0	0	0	1	1	1	0	0	0	0	Patient 2
43	1	0	0	0	0	0	1	0	0	1	1	1	0	0	0	1	0	0	1	0	1	0	1	0	0	Patient 3
27	0	1	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0	0	1	1	0	Patient 4
54	1	1	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	1	0	1	1	1	Patient N

FIGURE 3. Representation of the two-stage methodology. First stage: (a) patients represented by vectors of 2,265 features. They are the input to the (b) SOM (a grid of 25 nodes). The SOM output are 25 *prototype vectors* named as $\{v_i\}_{i=1}^{25}$. Second stage: prototype vectors are used as input for hierarchical clustering (c), providing 5 clusters which are depicted on the right part of the illustration. Nodes associated to each cluster appear with the same color.

of the input space through a neighbourhood function. This means that 'similar' samples from the input space will be assigned to neighbour nodes of the grid, enabling the possibility to discover groups of patients with common clinical characteristics and to visualize patterns of the population. Remark that the SOM mapping aids data visualization and may reveal hidden patterns from complex data in a straightforward way.

The basic SOM algorithm begins by assigning random values to the prototype vectors v_j, followed by an iterative training stage of a fixed number of iterations. For each iteration t, the next steps are performed:

- *Step 1.* A sample x_i is randomly chosen from X.
- *Step 2.* The sample x_i is compared with each prototype vector v_j by calculating the distance between them. The nearest node b to x_i is chosen as the winner node, called the Best Matching Unit (BMU). Formally, the BMU is defined as:

$$b = \operatorname*{argmin}_{j} \, d(x_i, v_j), \quad j \in 1, \cdots, M \tag{1}$$

where $d(.,.)$ indicates a distance metric.

- *Step 3.* The prototype vector associated to the BMU and those prototype vectors associated to the neighbouring nodes are updated according to the next learning rule:

$$v_j(t+1) = v_j(t) + \alpha(t) h_{jb}(t)(x_i(t) - v_j(t)), \tag{2}$$

for $j = 1, \cdots, M$, where t and $t+1$ are the current and the next iteration, respectively. We consider a monotically decreasing learning rate with time, given by $\alpha(t)$. The neighbourhood function $h_{jb}(t)$, which is centered on the winner node b, is given by $h_{jb}(t) = \exp\left(-\frac{\|r_b - r_j\|^2}{2\sigma^2(t)}\right)$, where r_b and r_j are the positions of nodes b and j on the grid, and σ is the neighborhood radius which decreases monotically with time $\sigma(t)$ [16], [17].

In our study, the prototype vector of each node represents a prototype patient. Each prototype patient characterizes individuals associated to that specific node based on clinical codes used (diagnosis and drug codes). Next, we propose to cluster prototype patients by leveraging the advantages the intrinsic similarity between these vectors. Note that neighbour nodes present a similarity relationship. This clustering provide new insights associated to patients with a similar chronic condition.

B. CLUSTERING OF THE PROTOTYPE VECTORS

Clustering aims to reveal underlying similarities and group samples based on distances, where smaller distances imply similar samples [61], [62]. Literature shows a variety of clustering methods [63], which are categorized into different types: partitional and hierarchical approaches [64]. Among partitional clustering, k-Means is the most popular. The number of K clusters is established a priori, where each cluster is represented by a centroid or cluster center. The centroid's location is found according to an iterative algorithm minimizing, for each cluster, the Euclidean distance between samples in the cluster and its corresponding centroid [65].

In contrast, in hierarchical clustering, samples are organized into a tree-like structure based on distance measures [66], [67]. The Agglomerative Hierarchical Clustering (AHC), which is the most common approach [68], starts considering each sample as a single cluster. At each iteration, the two clusters with smallest inter-cluster distance are merged into a new cluster. According to the inter-cluster distance, different algorithms can be found: *single linkage, average linkage, complete linkage and Ward* [67].

Following the approach adopted by various researchers [37], [69]–[72], we apply AHC on the prototype patients to provide an interpretable and visual characterization (tree-like structure) of the similarity between clusters.

To validate the cluster's quality and to determine an adequate number of clusters, we consider several clustering validity indices (CVIs) [73]–[76]. CVIs mainly consider two principles: (1) *compactness*, which measures how closely are the elements in the cluster (intra-cluster distance); and (ii) *separability* (inter-cluster distance) measuring the separation between clusters. Since clustering seeks to minimize intra-cluster distance (improving compactness) and to maximize inter-cluster distance (increasing the separation between clusters) [37], most of CVIs are based on compactness, separability or combination of both [77]. Though many CVIs have been proposed [73], [75], [78]–[80], this work considers the C index [79] and the Silhouette coefficient [81].

IV. EXPERIMENTS AND RESULTS

This section covers the experimental setup, the SOM visual interpretation, and the subsequent cluster characterization from a clinical viewpoint.

A. EXPERIMENTAL SETUP

In this work, the SOM is implemented using the Kohonen library [82] of the software environment R. The SOM training is performed using diagnosis and drug features. As stated in Section II, several parameters are involved in the SOM configuration. Thus, we explore different values for the number of iterations $\{1024, 2048, 3017, 4096\}$, for the learning rate values in the interval $[0.01, 0.05]$, and for the neighborhood radius σ in $[0.005, 0.3]$ Furthermore, we explore different structures for the two-dimensional grid based on the number of nodes. Specifically, we evaluated structures of 3×3 (9 nodes), 5×5 (25 nodes), 7×7 (49 nodes) and 10×10 (100 nodes). Our experiments showed that a two-dimensional grid of 5×5 was suitable to extract patterns of chronic diseases, to discover relationships among features, and to establish groups with specific clinical characteristics. As a result of the SOM training, a total of 25 prototype vectors were obtained. These vectors allow to capture intrinsic relationships from the data.

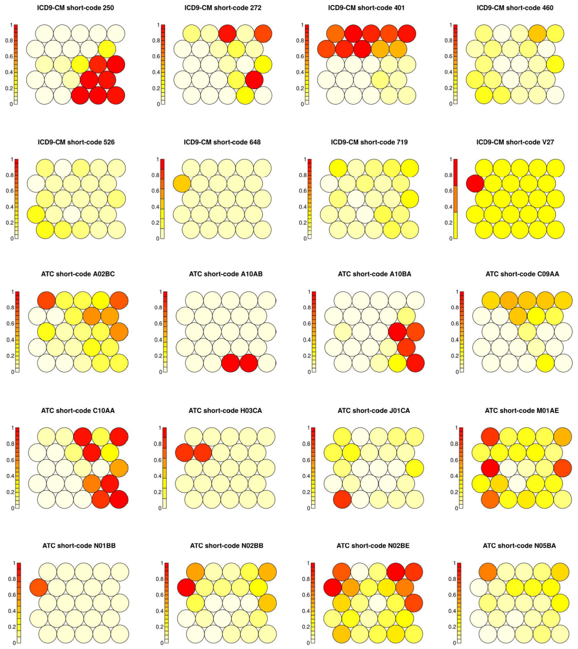


FIGURE 4. Selected CPRs of the SOM. CPRs for ICD9-CM codes: '250'; '272'; '401'; '460'; '526'; '648'; '719'; 'V27'. CPRs for ATC short-codes: 'A02BC'; 'A10AB'; 'A10BA'; 'C09AA'; 'C10AA'; 'H03CA'; 'J01CA'; 'M01AE'; 'N01BB'; 'N02BB'; 'N02BE'; 'N05BA'.

B. SOM VISUAL INTERPRETATION

The SOM has been applied for exploratory data analysis following a visual approach. Specifically, the *component plane representation* (CPR) has been broadly used for visualization and interpretation of high-dimensional data [83]–[85]. A CPR corresponds to a bi-dimensional grid of nodes (the same as the SOM architecture) showing the mapping of a feature of the prototype vectors on the grid. Each feature has its corresponding CPR, which is visualized by setting a color-coding. In our CPR visualizations, the darker the intensity color, the higher the mapping value of the corresponding feature.

These visualizations may reveal some associations between features by comparing two or more CPRs. This comparison is carried out considering positions and color-coding of the nodes. For example, if two CPRs show high color intensity in the same nodes of the grid, an association between both features in the corresponding prototypes can be assumed. In our case, we attempt to seek potential relationships among features linked to certain type of chronic patients according to diagnosis and drug codes. As we mentioned, we handled 2,265 features. Since showing all CPRs is not viable, we only show the CPRs of a determined set of features. Specifically, we considered those features with the highest average rate values in the diagnosis and drug profiles (see Fig. 2).

Starting with diabetic patients, we explore CPRs associated with ICD9-CM and ATC short-codes directly related to this chronic disease. The CPR of the ICD9-CM code '250' (first row, first column in Fig. 4) suggests that the majority of diabetic patients are distributed on the bottom right nodes.

The ATC short-codes most used in the treatment of diabetes are visually identified by ATC 'A10AB' (insulines) and 'A10BA' (biguanides). Regarding hypertensive patients, the CPR associated with ICD9-CM '401' (first row, third column) suggests that hypertensive patients are mainly located on the top nodes of the grid. By analyzing other CPRs, ATC 'C09AA' (best-known as statins) shows to be the main drug for hypertension treatment. CPR analysis also reveals the existence of ATC 'C10AA' (HMG CoA reductase inhibitors, fourth row, first column) and ICD9-CM '272' (Disorders of lipoid metabolism) (first row, second column) in certain nodes associated with diabetes and hypertension, which indicates that cholesterol and obesity are comorbidities present in patients diagnosed with these chronic conditions. Note also that ATC 'A02BC' (proton pump inhibitors (PPIs)) is common for the diabetes and hypertension drug profile as it can be shown in the CPR analysis (high intensity color at nodes linked to chronic diseases). Note that the increment of PPIs in relation to healthy patients is usually attached to the treatment of polymedicated patients, more frequent on elderly chronic patients. This increment, also evidenced by the SOM, has supported tools to reduce the PPIs prescription and therefore avoid their side effect on patients not directly benefited by these kind of drugs [86]. Some CPRs also evidence a moderate consumption of analgesics ('M01AE', fourth row, fourth column) and 'N02BE' (fifth row, third column) in nodes not only associated with chronic diseases, which is also evidenced comparing the peaks in the drug profiles of Fig. 2.

The CPR associated with ICD9-CM short-code 'V27' (Outcome of delivery, single liveborn, panel in second row, fourth column) was relevant to identify pregnant women. The node with the highest intensity in the color bar has also high intensity in the CPR '401' (hypertension). As a consequence, it may suggest a relation between pregnancy and hypertension. Nevertheless, these pregnant women did not take drugs for hypertension treatment. Another interesting finding is that most related CPRs were those identified to analgesics ATC 'N01BB', 'N02BB' and 'N02BE'. This makes sense, because on pregnancy status, prescription drugs are mainly analgesics. Furthermore, ICD9-CM short-code '648' (Other current conditions in the mother classifiable elsewhere but complicating pregnancy) is notorious since it is related to DM complicating pregnancy.

C. ESTIMATION OF NUMBER OF CLUSTERS

The insights gained by applying a SOM helped to identify patterns of chronic diseases. For this purpose, we propose to cluster prototype vectors linked to nodes aiming to discover groups of patients with shared chronic conditions. In order to compare different clustering approaches on the prototype vectors, we apply AHC with different *linkage* criteria (Ward, single, average and complete) and *k*-Means. We use the two CVIs presented in Section III to determine an appropriate number of clusters: (1) the Silhouette coefficient, and (2) the *C* index.

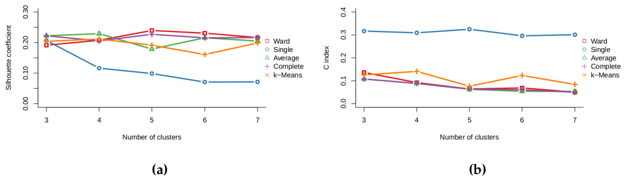


FIGURE 5. Cluster validity indices (CVIs) applied on prototype vectors. (a) Silhouette coefficient; (b) C index.

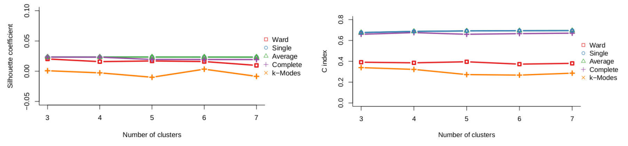
For a visual comparison, Fig. 5 shows the three CVIs considering different clustering methods and number of clusters. Note that the x-axis represents the number of clusters, while the y-axis shows the corresponding CVI value. Regarding the Silhouette coefficient (see Fig. 5 (a)), note that the Ward approach provides the highest value with five clusters. By analyzing the C index (see Fig. 5 (b)), where lower values represent a better clustering performance, Ward, average and complete linkage reach minimum values for both five and six clusters. Since different numbers of clusters can be a suitable election, a further exploratory analysis of the clusters reveals that the insights with five clusters are more clinically meaningful. In general, Ward demonstrates better clustering performance regarding CVIs values. Therefore, hereafter we explore the results of AHC Ward clustering with five clusters.

To complement the analysis, and to show the potential of the two-stage methodology, several clustering methods are carried out on the raw data. Specifically, we compare the results of *k*-Means and AHC with different linkage applied on the prototype vectors (Fig. 5) with the equivalent clustering methods on the raw data. Note that, in our case, the raw data are binary. Though a variety of clustering methods can be found in the literature, they are usually designed to work with numerical features, requiring some adaptations when clustering samples with categorical features. For instance, in AHC, distance measures that appropriately handle categorical features should be considered. In particular, the Jaccard distance has been used in this work [87]. Owing to the binary nature of the raw data features, *k*-Means is replaced by the *k*-Modes method, an extension of *k*-Means for categorical data [88].

The CVIs associated to *k*-Modes and AHC with different linkage on raw data are shown in Fig. 6. A brief inspection of Fig. 6 (a-b) shows lower values for the Silhouette coefficient in relation to those obtained when clustering prototype vectors: 0.025 versus 0.25 (see Fig. 5). Regarding the C index, best performance (lower C index) are obtained when clustering the prototype vectors than the raw data (0.09 versus 0.3). AHC Ward consistently provides the best performance for both CVIs when the prototype vectors are considered. From these outcomes, we can conclude that the two-stage methodology combining SOM and AHC Ward clustering is appropriate.

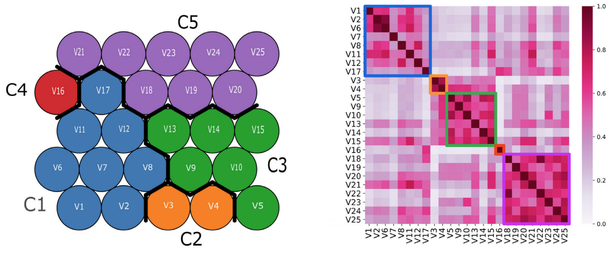
D. CLUSTER CHARACTERIZATION

In this section, a characterization of clusters using AHC Ward is accomplished through the most appropriated number of



(a) (b)

FIGURE 6. Cluster validity indices (CVIs) applied on raw data. (a) Silhouette coefficient; (b) C index.



(a) (b)

FIGURE 7. SOM grid and correlogram of the prototype vectors. (a) Grid of nodes and names of the prototype vectors, where each cluster is depicted by a different color (five clusters are considered); (b) Correlation coefficient matrix between pairs of prototype vectors, with axes indicating the prototype vectors according to (a).

clusters. Hereafter the analysis is carried out using five clusters, which are shown in Fig. 7 (a). Each cluster is represented by a different color: C1 (blue nodes), C2 (orange nodes), C3 (green nodes), C4 (red node), and C5 (purple nodes). In order to show the SOM topology-preserving property, we compute the Pearson correlation coefficient (PCC) between pairs of prototype vectors (see Fig. 7 (b) for details). The PCC is a numerical value ranging between $[-1, +1]$ which quantifies the linear relationship between two vectors [89], where 0 means no linear relationship, and $+1/-1$ represents an exact positive/negative linear relationship. For instance, taking the prototype vectors associated with nodes v_3 and v_4, we obtain a Pearson correlation coefficient around 0.8 (i.e. high linear correlation). This can be a reasonable result, since both prototypes are in the same cluster (cluster C2). In contrast, the value of this coefficient for the prototype vectors v_3 (cluster C2) and v_{17} (cluster C1) is around 0.1 (low linear correlation). Note that this happens because v_3 and v_{17} belong to different clusters and are spatially separated in the SOM grid.

To present these coefficients in a visual manner, a correlogram (a graph of a correlation matrix) is displayed in Fig. 7 (b). Note that prototype vectors do not appear with correlative numbers in rows and columns of this matrix. Instead, they have been arranged so that ‘similar’ prototypes are in adjacent positions, thus facilitating cluster visualization. Since prototypes in the same cluster are in adjacent positions, we have superimposed on the correlogram five squares with solid colors (the same as that of each cluster) to encompass prototypes in the same cluster.

The diagnosis and drug profiles associated with patients in each cluster are depicted in Fig. 8. These profiles provide a visual way to interpret the prevalence of different diagnosis and drug codes. The description of ICD9-CM and ATC short-codes linked to the highest values in the profiles are shown in Table 1. A detailed analysis of the obtained clusters is presented in the next paragraphs.

TABLE 1. Description of ICD9-CM and ATC short-codes with highest values in the diagnosis and drug profiles of Fig. 8.

ICD9-CM short-code	Description
‘250’	Diabetes mellitus
‘272’	Disorders of lipoid metabolism
‘401’	Essential hypertension
‘460’	Acute nasopharyngitis
‘526’	Diseases of the jaws
‘645’	Late pregnancy
‘648’	Diabetes mellitus of mother, complicating pregnancy
‘650’	Normal delivery
‘719’	Other and unspecified disorders of joint
‘V22’	Normal pregnancy
‘V27’	Outcome of delivery, single liveborn
ATC short-code	**Description**
‘A02BC’	Proton pump inhibitors
‘A10AB’	Insulins and analogues for injection, fast-acting
‘A10AE’	Insulins and analogues for injection, long-acting
‘A10BA’	Biguanides
‘C09AA’	ACE inhibitors plain
‘C10AA’	HMG CoA reductanse inhibitors
‘H03CA’	Iodine therapy
‘M01AE’	Propionic acid derivatives
‘N01BB’	Amides
‘N02BB’	Pyrazolones
‘N02BE’	Anilides
‘N05BA’	Benzodiazepine derivatives

Cluster C1 includes 2,437 individuals (mean age, 25.76 years), 81.45% associated to CRG-1000, 12.27% to CRG-5192 and 6.28% to CRG-5424. The mean age of patients in this cluster is the lowest of all clusters. None of the ICD9-CM/ATC short-codes of this cluster has a noteworthy average rate (see Fig. 8 (a)-(b)). An exploration of the drug profile (Fig. 8 (b)) shows that the most common drugs consumed by individuals of C1 are analgesics to treat general pain, where ‘N02BE’ and ‘M01AE’ are the ATC short-codes with highest average rate. Taking into account the percentages of individuals related to the majority class (CRG-1000) and the analysis of the profiles, we identify cluster C1 with the healthy population.

Cluster C2 contains 433 patients (mean age of 30.84 years), all of them linked to CRG-5424. From the diagnosis profile (Fig. 8 (c)), we conclude that the ICD9-CM short-code ‘250’ has the highest average rate. Other codes in the diagnosis profile (‘719’, ‘526’ and ‘460’) have a much lower average rate value, indicating that individuals in C2 are primarily diabetics. With regard to the drug profile (Fig. 8 (d)), ATC short-codes with the highest average rate are ‘A10AB’ and ‘A10AE’. Considering these results, we characterize patients in cluster C2 as young diabetics who consume mostly insulin.

Cluster C3 contains 1,753 patients (mean age, 55,58 years), 83.29% associated with CRG-5424, 9.12% with CRG-5192 and 7.59% with CRG-1000. For diagnoses

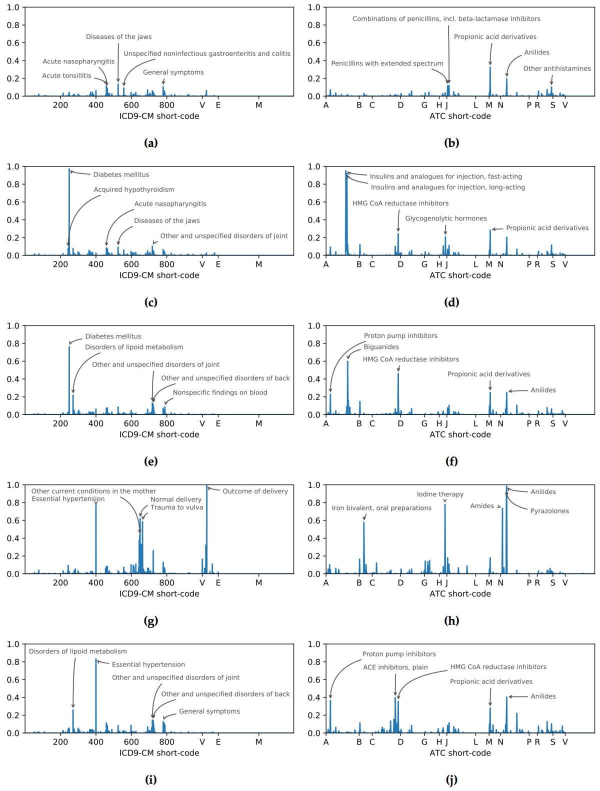


FIGURE 8. Diagnosis profiles (left panels) and drug profiles (right panels) for each of the five clusters: (a-b) C1; (c-d) C2; (e-f) C3; (g-h) C4; (i-j) C5.

(Fig. 8 (e)), the ICD9-CM short-codes with the highest average rate are ‘250’ and ‘272’. By exploring the drug profile in Fig. 6 (f), two ATC short-codes stand out above all: ‘A10BA’ and ‘C10AA’. ‘A10BA’ is used to treat non-insulin-dependant individuals [90], whereas ‘C10AA’ (best-known as statins) is used to reduce low-density lipoprotein cholesterol. Besides, it has properties that help for the prevention of cardiovascular events [91], [92]. In view of these insights, we characterize C3 as associated with the diabetic population. Note that the highest values of profiles of C3 correspond with the highest values of the profiles associated to CRG-5424 (Fig. 2 (e)-(f)). We conclude that the profiles associated with C3 are most similar to those of CRG-5424. Though clusters C2 and C3 are mainly related to the diabetic population, there is a remarkable difference in terms of drugs consumed. The individuals associated to C2 take biguanides and statins to a lesser extent.

Cluster C4 contains 145 patients (mean age, 30.25 years), 71.92% associated with CRG-5192, 17.93% with CRG-1000 and 10.34% with CRG-5424. All patients are pregnant women with diagnosis codes ‘645’, ‘648’, ‘650’, ‘V22’, ‘V27’ and ‘401’ (see the high values in Fig. 8 (g)). Note that most of the previous codes are strongly linked to pregnancy, with the exception of the ICD9-CM short-code ‘401’ (hypertension). Despite these pregnant women being coded with ‘401’, none of them takes drugs linked to hypertension such as angiotensin-converting enzyme (ACE) inhibitors or HMG CoA reductanse inhibitors. Concerning the drug consumption (Fig. 6 (h)), the most frequent ATC short-codes are ‘N02BB’ and ‘N02BE’. From a clinical viewpoint, it was validated that pregnant women are only allowed to take ‘N02BB’ and ‘N02BE’ (i.e. local anesthetics and analgesics). A previous study [93] showed that ACE inhibitors and angiotensin receptor blockers (ARBs) (drugs used to treat hypertensives) are fetotoxic and its discontinuation during pregnancy is highly recommended, due to malformations and adverse events that were reported [94].

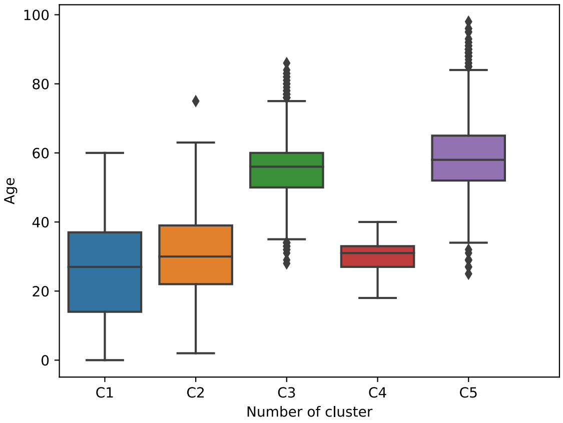


FIGURE 9. Boxplot of the age for each cluster.

Cluster C5 contains 1,730 patients (mean age, 55.74 years), 95.24% associated with CRG-5192, 3.84% with CRG-5424 and 0.92% with CRG-1000. The diagnosis profile (Fig. 8 (i)) shows that the ICD9-CM short-code ‘401’ has the highest average rate. For drugs (Fig. 6 (j)), the ATC short-codes with the highest average rate are ‘A02BC’, ‘C09AA’, ‘C10AA’. They are common drugs prescribed in the treatment of hypertension [95]. The drug therapy in hypertension aims to reduce the risks associated with high blood pressure (BP) [96]. This BP reduction requires a therapy with the combination of different drugs such as ACE inhibitors, ARBs, diuretics and β-blockers [96]. The use of several anti-hypertensives is evidenced in the drug profile, where the codes reach similar values of average rate and none of them predominates. By considering profiles linked to C5 and keeping in mind the majority class (CRG-5192), we associate this cluster with the hypertensive population.

Complementing the above cluster analysis, in Fig. 9 we provide the boxplot summarizing the distribution of the patient’s age for each cluster. Note that median age of clusters C1, C2 and C4 is significantly lower than that of C3 and C5, which are mainly associated with chronic patients.

E. CLINICAL VALIDATION USING CRGs

In order to validate clusters found with the two-stage methodology from a clinical viewpoint, we quantify the relationship between the clusters profiles (see Fig. 8) and CRGs profiles (see Fig. 2) by means of the Pearson correlation coefficient (PCCs). Thus, we compute the PCC between the

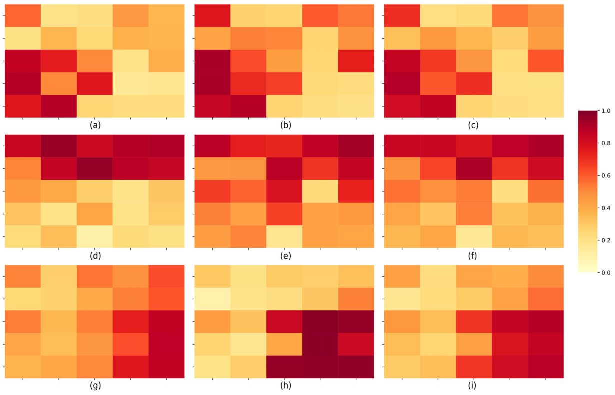


FIGURE 10. LCM representations. LCM between the profile of each node and the profile of each CRG: (a-c) for CRG-1000; (d-f) for CRG-5192; (g-i) for CRG-5424. (Left panels) diagnosis profiles; (middle panels) drug profiles; (right panels) concatenation of diagnosis and drug profiles.

diagnosis/drug profile associated with each CRG and the profile computed for: (1) each node, and (2) each cluster. We firstly obtain the three types of profiles linked to each node (diagnosis, drug and concatenation of both) and we compute the PCC between each one and the profile corresponding to each CRG. Bearing in mind the grid size (i.e. 25 nodes), we get a total of 25 PCC values per type of profile associated to one CRG. In order to have a visual way for interpreting these results, we propose to show PCC values as a heatmap on a bi-dimensional array of size 5 × 5 representing the SOM nodes. The heatmap representation uses different color intensities to illustrate the highest and lowest PCC values, providing what we have named a 'linear correlation map' (LCM). Each rectangle of the LCM corresponds to one specific node of the grid. Taking into account that we handled three types of profiles, it results in three LCMs by CRG, reaching a total of nine LCMs for the analysis (see Fig. 10). The first row of the panels (a-c in Fig. 10) refers to the three LCMs associated with CRG-1000; the second row of panels (d-f in Fig. 10) shows LCMs corresponding to CRG-5192, and the last row (g-i in Fig. 10) presents LCMs linked to CRG-5424. Note that LCM can contribute to characterize each node from a quantitative and clinical viewpoint by identifying which nodes are closely related a chronic diseases.

From Fig. 10, it becomes straightforward to distinguish the predominant health-status associated to each node by identifying the most intense red color in the bi-dimensional heatmap. Taking into account the LCMs in Fig. 10 (a-c), the highest PCC values are at the bottom left nodes, indicating that healthy patients are mostly located on those nodes. Considering Fig. 10 (d) (i.e. PCC between diagnosis profiles of CRG-5192 and these associated to each node), high values of correlation are placed in the top nodes, corresponding to a hypertensive population. The highest PCC values on the LCM for drugs (see Fig. 10 (e)) are not just on the same nodes as in Fig. 10 (d), but there are more nodes with high values. This goes in hand with the findings provided by the drug profile associated with CRG-5192 (see Fig. 2 (d)), where the highest peaks have quite similar values and many of them are related to analgesics (common drugs used in other CRGs, specially in healthy populations). Note that Fig. 10 (f) is a contribution of the two aforementioned LCMs. By comparing LCMs in the third row, it can be seen that the red color is more intense in nodes at the bottom right of the grid, for both diagnosis and drug profiles (see Fig. 10 (g-h)). Panel (i) shows the contribution of these two LCMs. Remark that drugs allow identification of nodes associated with diabetic patients in a better way than diagnoses. When comparing the LCMs associated with CRG-5192 and CRG-5424, it is evidenced that the drug therapy for DM is well-established (mainly insulin and biguanides), while the spectrum of drugs for hypertension is wider. Besides, it is frequent that common drugs for hypertension are also used for cardiovascular conditions.

TABLE 2. PCC values between profiles of each cluster (from C1 to C5) and those associated with: CRG-1000 (first row), CRG-5192 (second row), and CRG-5424 (third row). For each cluster, the last row shows the percentage of patients by CRG.

	C1	C2	C3	C4	C5
PCC between cluster profiles and **CRG-1000** diagnosis profiles **CRG-1000** drug profiles	 0.973 0.992	 0.250 0.265	 0.275 0.434	 0.197 0.436	 0.354 0.593
PCC between cluster profiles and **CRG-5192** diagnosis profiles **CRG-5192** drug profiles	 0.505 0.674	 0.151 0.272	 0.276 0.598	 0.511 0.462	 0.991 0.990
PCC between cluster profiles and **CRG-5424** diagnosis profiles **CRG-5424** drug profiles	 0.388 0.513	 0.978 0.618	 0.996 0.953	 0.100 0.260	 0.314 0.609
Patients (%) associated to: CRG-1000 CRG-5192 CRG-5424	 81.45 12.27 6.28	 0.0 0.0 100.0	 7.59 9.12 83.29	 17.93 71.72 10.34	 0.92 95.24 3.84

We expanded the analysis of correlation and ascertain whether the clusters characterized in Subsection IV-D (see Fig. 8) have a clinical interpretation in terms of PCCs. Towards that end, we first computed the diagnosis and drug profiles associated with each cluster. Then, the PCC between these profiles and those linked to the CRGs were evaluated. The resulting PCC values are shown in Table 2. Furthermore, for each cluster, we show the percentage of patients associated with each CRG. Note that cluster C1 was highly related to CRG-1000 in the two profiles (PCC values $>$ 0.9). Clusters C2 and C3, which have as majority class the CRG-5424, present a high relationship with the diagnosis profile of CRG-5424. However, the drug profile of C2 showed a low linear relationship with that associated with CRG-5424 (PCC value $\approx$ 0.6). The reason can be that the most common treatment in C2 is insulin (see Fig. 8 (d)), which is not matching with the highest values in the drug profile of CRG-5424 (see Fig. 2 (f)). The majority of patients in clusters C4 and C5 are labelled as hypertensive (CRG-5192). However, the profiles of C4 did not present high correlation with any CRGs. On the contrary, cluster C5 is closely related to the drug and diagnosis profiles of CRG-5192 (PCC values $>$ 0.9).

V. DISCUSSION

In this section we highlighted the main insights obtained from this work. We have shown that the two-stage approach allowed us to identify two kinds of diabetic patients in CRG-5424: insulin-dependant (cluster C2) and non-insulin-dependant (cluster C3). As presented in the drug profiles of Fig. 8, individuals in cluster C2 took mostly insulins (ATC short-codes ‘A10AB’ and ‘A10AE’), while those in cluster C3 were mainly characterized by consuming biguanides (ATC short-code ‘A10BA’), which is also a common drug for diabetes treatment (see the CRG-5424 profile in Fig. 2 (f)). The comparison of the aforementioned profiles was helpful to distinguish major differences between individuals in C2 and C3 in relation to the drug consumption rate. Additionally, the mean age of individuals in clusters C2 and C3 also pointed out a distinction between populations, identifying adult and elderly diabetics.

We also highlighted the presence of a specific cluster (C4) just associated to hypertensive women. In Fig. 8 (g), the highest average rate was associated with ICD9-CM ‘V27’, indicating that most individuals in this cluster had a single live birth. Furthermore, they mostly have the ICD9-CM code ‘401’ (Essential hypertension), and the code ‘648’ (Other current conditions in the mother complicating pregnancy childbirth) is also present to a lesser extent. In the literature review, we found that one of the major pregnancy complications is the onset of hypertension and diabetes [97], [98]. In fact, a high percentage of pregnant women may develop high blood pressure, that in certain cases may provoke acute conditions such as preeclampsia [99]. Some drugs used in hypertension and diabetes therapy (ACE inhibitors, ARBs, biguanides) are fetotoxic due to malformations and adverse events [93], [94], [100], and its discontinuation during pregnancy is highly recommended. The absence of drugs linked to hypertension (ACE inhibitors, ARBs, diuretics) and diabetes (biguanides) is evident from C4. The analysis of CPRs showed the prevalence of drugs ‘H03CA’, ‘N01BB’, ‘N02BB’, ‘N02BE’, which is coherent with drugs prescribed for pregnant women. The existence of cluster C4 certainly represented an unexpected finding from our analysis.

From the aforementioned outcomes, we confirmed that our data-driven two-stage methodology has potential to discover novel groups of patients in an unsupervised way. Our methodology raises as a suitable alternative for building appropriate clinical groups, becoming a promising approach with potential to be applied to any population due to its data-driven philosophy.

Our methodology primarily focuses on two chronic conditions (diabetes and hypertension), demonstrating notable outcomes for characterizing patient health-status addressing high-dimensionality issues. This work can be extrapolated to other scenarios with complex chronic diseases, for instance, patients suffering from multimorbidity (two or more chronic conditions at same time). Experiments with multimorbidity is out of the scope of the paper, and it has been considered as future work.

VI. CONCLUSION

The SOM has proven to be effective to find patterns in data recorded in the EHR. It showed great ability to visualize high-dimensional clinical data and demonstrated to be a powerful visual tool for finding patterns related to chronic diseases. The prototype vectors characterizing the nodes of the SOM grid are relevant because they capture intrinsic relationships from data while keeping the topology-preserving property. The proposed two-stage methodology (composed of the SOM and AHC) worked reasonably well to characterize chronic individuals and to distinguish clusters with discriminant characteristics from a clinical viewpoint. It could also be considered as a chance to conduct research into clinical data of specific groups of patients and, maybe, find novel co-factors (diseases or drugs) which could result in a different evolution of their health status. The characterization carried out by the profiles and CPRs provided a better understanding of chronic conditions. Finally, this characterization could support clinical decisions, impacting on the evolution of the chronic condition or the health-status. The positive impact of these decisions on cost reduction and patient’s satisfaction is remarkable both from a clinical and socioeconomic perspective.

ACKNOWLEDGMENT

This paper resulted from the efforts of many people of the clinical sphere, including Angel Gómez, Isabel Caballero, Javier Rivas, and José Luis Rodríguez from the University Hospital of Sureste, Madrid, Spain, and the University Hospital of Fuenlabrada, Madrid, Spain. David Chushig-Muzo would like to thank Wouter Labuschagne, Marissa Labuschagne, Nuno da Costa, and Frederick Atiah for the support and assistance during his research stay in South Africa.

AUTHOR CONTRIBUTIONS

DCM, CSR, IMJ wrote the manuscript. AE helped with the design of methods, reviewed and contributed for improving the manuscript. DCM implemented the methods, conducted statistical analyses. PMB guided the data acquisition and helped with clinical interpretation. CSR and IMJ contributed in methodology, data cleaning and pre-processing. All authors reviewed and approved the final version for submission.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] *Global Report on Diabetes: World Health Organization*, World Health Org., Geneva, Switzerland, 2016.

[2] *A Global Brief on Hypertension: Silent Killer, Global Public Health Crisis: World Health Day 2013*, World Health Org., Geneva, Switzerland, 2013.

[3] D. Yach, C. Hawkes, C. L. Gould, and K. J. Hofman, “The global burden of chronic diseases overcoming impediments to prevention and control,” *JAMA*, vol. 291, no. 21, pp. 2616–2622, 2004.

[4] L. M. Renard, V. Bocquet, G. Vidal-Trecan, M.-L. Lair, S. Couffignal, and C. Blum-Boisgard, "An algorithm to identify patients with treated type 2 diabetes using medico-administrative data," *BMC Med. Informat. Decis. Making*, vol. 11, no. 1, p. 23, Dec. 2011.

[5] W.-J. Guan, M. Jiang, Y.-H. Gao, H.-M. Li, G. Xu, J.-P. Zheng, R.-C. Chen, and N.-S. Zhong, "Unsupervised learning technique identifies bronchiectasis phenotypes with distinct clinical characteristics," *Int. J. Tuberculosis Lung Disease*, vol. 20, no. 3, pp. 402–410, Mar. 2016.

[6] B. K. Beaulieu-Jones and C. S. Greene, "Semi-supervised learning of the electronic health record for phenotype stratification," *J. Biomed. Informat.*, vol. 64, pp. 168–178, Dec. 2016.

[7] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," *J. Amer. Med. Informat. Assoc.*, vol. 24, no. 2, pp. 361–370, Mar. 2017.

[8] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018.

[9] C. Soguero-Ruiz, I. Mora-Jiménez, M. A. Mohedano-Munoz, M. Rubio-Sanchez, P. D. Miguel-Bohoyo, and A. Sanchez, "Visually guided classification trees for analyzing chronic patients," *BMC Bioinf.*, vol. 21, no. S2, pp. 1–19, Mar. 2020.

[10] S. R. Michalski, G. J. Carbonell, and M. T. Mitchell, Eds., *Machine Learning an Artificial Intelligence Approach*, vol. 2. San Francisco, CA, USA: Morgan Kaufmann, 1986.

[11] K. M. Brelsford, S. E. Spratt, and L. M. Beskow, "Research use of electronic health records: Patients' perspectives on contact by researchers," *J. Amer. Med. Inform. Assoc.*, vol. 25, no. 9, pp. 1122–1129, Sep. 2018.

[12] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, and H. Liu, "Clinical information extraction applications: A literature review," *J. Biomed. Informat.*, vol. 77, pp. 34–49, Jan. 2018.

[13] Z. Wang, A. D. Shah, A. R. Tate, S. Denaxas, J. Shawe-Taylor, and H. Hemingway, "Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning," *PLoS ONE*, vol. 7, no. 1, Jan. 2012, Art. no. e30412.

[14] M. Zhu, J. Xia, X. Jin, M. Yan, G. Cai, J. Yan, and G. Ning, "Class weights random forest algorithm for processing class imbalanced medical data," *IEEE Access*, vol. 6, pp. 4641–4652, 2018.

[15] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A machine learning methodology for diagnosing chronic kidney disease," *IEEE Access*, vol. 8, pp. 20991–21002, 2020.

[16] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biol. Cybern.*, vol. 43, no. 1, pp. 59–69, 1982.

[17] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, no. 9, pp. 1464–1480, Sep. 1990.

[18] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, nos. 1–3, pp. 1–6, 1998.

[19] C. Lin, C.-M. Lin, S.-T. Li, and S.-C. Kuo, "Intelligent physician segmentation and management based on KDD approach," *Expert Syst. Appl.*, vol. 34, no. 3, pp. 1963–1973, Apr. 2008.

[20] M. Pisati, C. T. Whelan, M. Lucchini, and B. Maître, "Mapping patterns of multiple deprivation using self-organising maps: An application to EU-SILC data for Ireland," *Social Sci. Res.*, vol. 39, no. 3, pp. 405–418, May 2010.

[21] K.-C. Hsu and S.-T. Li, "Clustering spatial–temporal precipitation data using wavelet transform and self-organizing map neural network," *Adv. Water Resour.*, vol. 33, no. 2, pp. 190–200, Feb. 2010.

[22] P. Louis, A. Seret, and B. Baesens, "Financial efficiency and social impact of microfinance institutions using self-organizing maps," *World Develop.*, vol. 46, pp. 197–210, Jun. 2013.

[23] M. Peleg, N. Asbeh, T. Kuflik, and M. Schertz, "Onto-clust—A methodology for combining clustering analysis and ontological methods for identifying groups of comorbidities for developmental disorders," *J. Biomed. Informat.*, vol. 42, no. 1, pp. 165–175, Feb. 2009.

[24] H.-C. Chou, C.-H. Cheng, and J.-R. Chang, "Extracting drug utilization knowledge using self-organizing map and rough set theory," *Expert Syst. Appl.*, vol. 33, no. 2, pp. 499–508, Aug. 2007.

[25] T. Faisal, F. Ibrahim, and M. N. Taib, "A noninvasive intelligent approach for predicting the risk in dengue patients," *Expert Syst. Appl.*, vol. 37, no. 3, pp. 2175–2181, Mar. 2010.

[26] M. Xu, T. C. Wong, and K. S. Chin, "A medical procedure-based patient grouping method for an emergency department," *Appl. Soft Comput.*, vol. 14, pp. 31–37, Jan. 2014.

[27] M. K. Markey, J. Y. Lo, G. D. Tourassi, and C. E. Floyd, Jr., "Self-organizing map for cluster analysis of a breast cancer database," *Artif. Intell. Med.*, vol. 27, no. 2, pp. 113–127, Feb. 2003.

[28] J. J. Liszka-Hackzell and D. P. Martin, "Analysis of nighttime activity and daytime pain in patients with chronic back pain using a self-organizing map neural network," *J. Clin. Monitor. Comput.*, vol. 19, no. 6, pp. 411–414, Dec. 2005.

[29] D.-R. Chen, R.-F. Chang, and Y.-L. Huang, "Breast cancer diagnosis using self-organizing map for sonography," *Ultrasound Med. Biol.*, vol. 26, no. 3, pp. 405–411, Mar. 2000.

[30] M. Schmuker, F. Schwarte, A. Brück, E. Proschak, Y. Tanrikulu, A. Givehchi, K. Scheiffele, and G. Schneider, "Sommer: Self-organising maps for education and research," *J. Mol. Model.*, vol. 13, no. 1, pp. 225–228, 2007.

[31] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," *Decis. Support Syst.*, vol. 50, no. 3, pp. 559–569, Feb. 2011.

[32] A. Astel, S. Tsakovski, P. Barbieri, and V. Simeonov, "Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets," *Water Res.*, vol. 41, no. 19, pp. 4566–4578, Nov. 2007.

[33] M. J. Cohen, A. D. Grossman, D. Morabito, M. M. Knudson, A. J. Butte, and G. T. Manley, "Identification of complex metabolic states in critically injured patients using bioinformatic cluster analysis," *Crit. Care*, vol. 14, no. 1, p. R10, 2010.

[34] Z. Afzal, M. J. Schuemie, J. C. van Blijderveen, E. F. Sen, M. C. Sturkenboom, and J. A. Kors, "Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records," *BMC Med. Informat. Decis. Making*, vol. 13, no. 1, p. 30, 2013.

[35] N. Liu, Z. X. Koh, J. Goh, Z. Lin, B. Haaland, B. P. Ting, and M. E. H. Ong, "Prediction of adverse cardiac events in emergency department patients with chest pain using machine learning for variable selection," *BMC Med. Informat. Decis. Making*, vol. 14, no. 1, p. 75, Dec. 2014.

[36] S. Van Gassen, B. Callebaut, M. J. Van Helden, B. N. Lambrecht, P. Demeester, T. Dhaene, and Y. Saeys, "FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data," *Cytometry Part A*, vol. 87, no. 7, pp. 636–645, Jul. 2015.

[37] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Trans. Neural Netw.*, vol. 11, no. 3, pp. 586–600, May 2000.

[38] *International Classification of Diseases, 9th Revision, Clinical Modification: Physician ICD-9-CM, 2005: Color-Coded, Illustrated*, Amer. Med. Assoc., Chicago, IL, USA, 2004, vols. 1–2.

[39] *The Anatomical Therapeutic Chemical Classification System With Defined Daily Doses (ATC/DDD)*, WHO, Oslo, Norway, 2006.

[40] A. A. Ginde, P. G. Blanc, R. M. Lieberman, and C. A. Camargo, "Validation of ICD-9-CM coding algorithm for improved identification of hypoglycemia visits," *BMC Endocrine Disorders*, vol. 8, no. 1, p. 4, 2008.

[41] C. R. Cooke, M. J. Joo, S. M. Anderson, T. A. Lee, E. M. Udris, E. Johnson, and D. H. Au, "The validity of using ICD-9 codes and pharmacy records to identify patients with chronic obstructive pulmonary disease," *BMC Health Services Res.*, vol. 11, no. 1, p. 37, Dec. 2011.

[42] M. Osler, S. Mårtensson, I. K. Wium-Andersen, E. Prescott, P. K. Andersen, T. S. H. Jørgensen, K. Carlsen, M. K. Wium-Andersen, and M. B. Jørgensen, "Depression after first hospital admission for acute coronary syndrome: A study of time of onset and impact on survival," *Amer. J. Epidemiol.*, vol. 183, no. 3, pp. 218–226, Feb. 2016.

[43] C. Bouza, T. Lopez-Cuadrado, and J. M. Amate-Blanco, "Use of explicit ICD9-CM codes to identify adult severe sepsis: Impacts on epidemiological estimates," *Crit. Care*, vol. 20, no. 1, p. 313, Dec. 2016.

[44] C. Soguero-Ruiz, A. A. Díaz-Plaza, P. de Miguel Bohoyo, J. Ramos-López, M. Rubio-Sánchez, A. Sánchez, and I. Mora-Jiménez, "On the use of decision trees based on diagnosis and drug codes for analyzing chronic patients," in *Proc. Int. Conf. Bioinf. Biomed. Eng.* Cham, Switzerland: Springer, 2018, pp. 135–148.

[45] A. Sánchez, C. Soguero-Ruiz, I. Mora-Jiménez, F. J. Rivas-Flores, D. J. Lehmann, and M. Rubio-Sánchez, "Scaled radial axes for interactive visual feature selection: A case study for analyzing chronic conditions," *Expert Syst. Appl.*, vol. 100, pp. 182–196, Jun. 2018.

[46] G. Chen, N. Khan, R. Walker, and H. Quan, "Validating ICD coding algorithms for diabetes mellitus from administrative data," *Diabetes Res. Clin. Pract.*, vol. 89, no. 2, pp. 189–195, Aug. 2010.

[47] J. C. Zgibor, T. J. Orchard, M. Saul, G. Piatt, K. Ruppert, A. Stewart, and L. M. Siminerio, "Developing and validating a diabetes database in a large health system," *Diabetes Res. Clin. Pract.*, vol. 75, no. 3, pp. 313–319, Mar. 2007.

[48] J. Huang, C. Osorio, and L. W. Sy, "An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes," *Comput. Methods Programs Biomed.*, vol. 177, pp. 141–153, Aug. 2019.

[49] K. Malloch and A. Conovaloff, "Patient classification systems, part 1: The third generation," *J. Nursing Admin.*, vol. 29, nos. 7–8, pp. 49–56, 1999.

[50] J. S. Hughes, R. F. Averill, J. Eisenhandler, N. I. Goldfield, J. Muldoon, J. M. Neff, and J. C. Gay, "Clinical risk groups (CRGs): A classification system for risk-adjusted capitation-based payment and health care management," *Med. Care*, vol. 42, no. 1, pp. 81–90, Jan. 2004.

[51] J. M. Neff, V. L. Sharp, J. Muldoon, J. Graham, J. Popalisky, and J. C. Gay, "Identifying and classifying children with chronic conditions using administrative data with the clinical risk group classification system," *Ambulatory Pediatrics*, vol. 2, no. 1, pp. 71–79, Jan. 2002.

[52] J. G. Berry, M. Hall, D. E. Hall, D. Z. Kuo, E. Cohen, R. Agrawal, K. D. Mandl, H. Clifton, and J. Neff, "Inpatient growth and resource use in 28 children's hospitals: A longitudinal, multi-institutional study," *JAMA Pediatrics*, vol. 167, no. 2, pp. 170–177, 2013.

[53] P. D. Hain, J. C. Gay, T. W. Berutti, G. M. Whitney, W. Wang, and B. R. Saville, "Preventability of early readmissions at a children's hospital," *Pediatrics*, vol. 131, no. 1, pp. e171–e181, 2013.

[54] J. Fernández-Sánchez, C. Soguero-Ruiz, P. de Miguel-Bohoyo, F. J. Rivas-Flores, A. Gómez-Delgado, F. J. Gutiérrez-Expósito, and I. Mora-Jiménez, "Clinical risk groups analysis for chronic hypertensive patients in terms of ICD9-CM diagnosis codes," in *Proc. 4th Int. Conf. Physiol. Comput. Syst.*, Madrid, Spain. Setúbal, Portugal: SciTePress, vol. 1, 2017, pp. 13–22.

[55] T.-M. Chan, Y. Li, C.-C. Chiau, J. Zhu, J. Jiang, and Y. Huo, "Imbalanced target prediction with pattern discovery on clinical data repositories," *BMC Med. Informat. Decis. Making*, vol. 17, no. 1, p. 47, Dec. 2017.

[56] A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, A. Hawalah, and A. Hussain, "Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study," *IEEE Access*, vol. 4, pp. 7940–7957, 2016.

[57] S. Huda, K. Liu, M. Abdelrazek, A. Ibrahim, S. Alyahya, H. Al-Dossari, and S. Ahmad, "An ensemble oversampling model for class imbalance problem in software defect prediction," *IEEE Access*, vol. 6, pp. 24184–24195, 2018.

[58] S. H. Ebenuwa, M. S. Sharif, M. Alazab, and A. Al-Nemrat, "Variance ranking attributes selection techniques for binary classification problem in imbalance data," *IEEE Access*, vol. 7, pp. 24649–24666, 2019.

[59] T. R. Hoens and N. V. Chawla, "Imbalanced datasets: From sampling to classifiers," in *Proc. Imbalanced Learn., Found., Algorithms, Appl.*, 2013, pp. 43–59.

[60] S. Turcan *et al.*, "IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype," *Nature*, vol. 483, no. 7390, p. 479, 2012.

[61] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). Berlin, Germany: Springer-Verlag, 2006.

[62] F. Cao, J. Liang, D. Li, L. Bai, and C. Dang, "A dissimilarity measure for the k-Modes clustering algorithm," *Knowl.-Based Syst.*, vol. 26, pp. 120–127, Feb. 2012.

[63] M. G. H. Omran, A. P. Engelbrecht, and A. Salman, "An overview of clustering methods," *Intell. Data Anal.*, vol. 11, no. 6, pp. 583–605, Nov. 2007.

[64] R. Cordeiro de Amorim and B. Mirkin, "Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering," *Pattern Recognit.*, vol. 45, no. 3, pp. 1061–1075, Mar. 2012.

[65] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, Oakland, CA, USA, vol. 1, 1967, pp. 281–297.

[66] R. Xu and D. Wunsch, II, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

[67] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, Dec. 2017.

[68] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: An overview," *WIREs Data Mining Knowl. Discovery*, vol. 2, no. 1, pp. 86–97, Jan. 2012.

[69] J. E. Calamari, P. S. Wiegartz, and A. S. Janeck, "Obsessive–compulsive disorder subgroups: A symptom-based clustering approach," *Behaviour Res. Therapy*, vol. 37, no. 2, pp. 113–125, Feb. 1999.

[70] C. Ambroise, G. Sèze, F. Badran, and S. Thiria, "Hierarchical clustering of self-organizing maps for cloud classification," *Neurocomputing*, vol. 30, nos. 1–4, pp. 47–52, Jan. 2000.

[71] Y.-S. Park, J. Tison, S. Lek, J.-L. Giraudel, M. Coste, and F. Delmas, "Application of a self-organizing map to select representative species in multivariate analysis: A case study determining diatom distribution patterns across France," *Ecological Informat.*, vol. 1, no. 3, pp. 247–257, Nov. 2006.

[72] M. L. Gonçalves, M. L. A. Netto, J. A. F. Costa, and J. Z. Júnior, "An unsupervised method of classifying remotely sensed images using kohonen self-organizing maps and agglomerative hierarchical clustering methods," *Int. J. Remote Sens.*, vol. 29, no. 11, pp. 3171–3207, Jun. 2008.

[73] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, Jun. 1985.

[74] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *J. Cybern.*, vol. 4, no. 1, pp. 95–104, Jan. 1974.

[75] M. Brun, C. Sima, J. Hua, J. Lowey, B. Carroll, E. Suh, and E. R. Dougherty, "Model-based evaluation of clustering validation measures," *Pattern Recognit.*, vol. 40, no. 3, pp. 807–824, Mar. 2007.

[76] J. Handl, J. Knowles, and D. B. Kell, "Computational cluster validation in post-genomic data analysis," *Bioinformatics*, vol. 21, no. 15, pp. 3201–3212, Aug. 2005.

[77] M. Kim and R. S. Ramakrishna, "New indices for cluster validity assessment," *Pattern Recognit. Lett.*, vol. 26, no. 15, pp. 2353–2363, Nov. 2005.

[78] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1650–1654, Dec. 2002.

[79] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognit.*, vol. 46, no. 1, pp. 243–256, Jan. 2013.

[80] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *J. Intell. Inf. Syst.*, vol. 17, nos. 2–3, pp. 107–145, Dec. 2001.

[81] R. C. de Amorim and C. Hennig, "Recovering the number of clusters in data sets with noise features using feature rescaling factors," *Inf. Sci.*, vol. 324, pp. 126–145, Dec. 2015.

[82] R. Wehrens and L. M. C. Buydens, "Self-and super-organizing maps in R: The Kohonen package," *J. Stat. Softw.*, vol. 21, no. 5, pp. 1–19, 2007.

[83] J. Vesanto, "SOM-based data visualization methods," *Intell. Data Anal.*, vol. 3, no. 2, pp. 111–126, Mar. 1999.

[84] M. Alvarez-Guerra, C. González-Piñuela, A. Andrés, B. Galán, and J. R. Viguri, "Assessment of self-organizing map artificial neural networks for the classification of sediment quality," *Environ. Int.*, vol. 34, no. 6, pp. 782–790, Aug. 2008.

[85] J. Huysmans, B. Baesens, J. Vanthienen, and T. van Gestel, "Failure prediction with self organizing maps," *Expert Syst. Appl.*, vol. 30, no. 3, pp. 479–487, Apr. 2006.

[86] A. Rieckert, A. Becker, N. Donner-Banzhof, A. Viniol, B. Bücker, S. Wilm, A. Sönnichsen, and A. Barzel, "Reduction of the long-term use of proton pump inhibitors by a patient-oriented electronic decision support tool (arriba-PPI): Study protocol for a randomized controlled trial," *Trials*, vol. 20, no. 1, p. 636, Dec. 2019.

[87] Z. Huang, "Extensions to the K-means algorithm for clustering large data sets with categorical values," *Data Mining Knowl. Discovery*, vol. 2, no. 3, pp. 283–304, Sep. 1998.

[88] A. Chaturvedi, P. E. Green, and J. D. Caroll, "K-modes clustering," *J. Classification*, vol. 18, no. 1, pp. 35–55, Jan. 2001.

[89] J. Lee Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *Amer. Statistician*, vol. 42, no. 1, pp. 59–66, Feb. 1988.

[90] C. J. Bailey, "Biguanides and NIDDM," *Diabetes Care*, vol. 15, no. 6, pp. 755–772, Jun. 1992.

[91] N. L. Zaharan, D. Williams, and K. Bennett, "Statins and risk of treated incident diabetes in a primary care population," *Brit. J. Clin. Pharmacol.*, vol. 75, no. 4, pp. 1118–1124, Apr. 2013.

[92] A. A. Carter, T. Gomes, X. Camacho, D. N. Juurlink, B. R. Shah, and M. M. Mamdani, "Risk of incident diabetes among patients treated with statins: Population based study," *BMJ*, vol. 346, 2013.

[93] P. R. James and C. Nelson-Piercy, "Management of hypertension before, during, and after pregnancy," *Heart*, vol. 90, no. 12, pp. 1499–1504, Dec. 2004.

[94] K. Della-Giustina and G. Chow, "Medications in pregnancy and lactation," *Emergency Med. clinics North Amer.*, vol. 21, no. 3, pp. 585–613, 2003.

[95] E. P. K. Woolthuis, W. J. de Grauw, W. H. van Gerwen, H. J. van den Hoogen, E. H. van de Lisdonk, J. F. Metsemakers, and C. van Weel, "Identifying people at risk for undiagnosed type 2 diabetes using the GP's electronic medical record," *Family Pract.*, vol. 24, no. 3, pp. 230–236, May 2007.

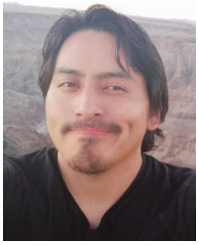
[96] A. H. Gradman, J. N. Basile, B. L. Carter, and G. L. Bakris, "Combination therapy in hypertension," *J. Amer. Soc. Hypertension*, vol. 4, no. 2, pp. 90–98, 2010.

[97] P. Saudan, M. A. Brown, M. L. Buddle, and M. Jones, "Does gestational hypertension become pre-eclampsia?" *BJOG: Int. J. Obstetrics Gynaecol.*, vol. 105, no. 11, pp. 1177–1184, Nov. 1998.

[98] L. Bellamy, J.-P. Casas, A. D. Hingorani, and D. Williams, "Type 2 diabetes mellitus after gestational diabetes: A systematic review and meta-analysis," *Lancet*, vol. 373, no. 9677, pp. 1773–1779, May 2009.

[99] B. Sibai, "Diagnosis and management of gestational hypertension and preeclampsia," *Obstetrics Gynecology*, vol. 102, no. 1, pp. 181–192, Jul. 2003.

[100] R. S. Lindsay and M. R. Loeken, "Metformin use in pregnancy: Promises and uncertainties," *Diabetologia*, vol. 60, no. 9, pp. 1612–1619, Sep. 2017.



DAVID CHUSHIG-MUZO received the B.Sc. and M.Sc. degrees in telecommunications engineering from Rey Juan Carlos University. He is currently pursuing the Ph.D. degree in machine learning with the Department of Signal Theory and Communications, Telematics and Computing. His main research interests include statistical learning theory, machine learning, data mining, and analytics.

CRISTINA SOGUERO-RUIZ received the Ph.D. degree in machine learning with applications in healthcare, with the Joint Doctoral Program in Multimedia and Communications in conjunction with Rey Juan Carlos University and the University Carlos III of Madrid, in 2015. She was supported by the FPU Spanish Research and Teaching Fellowship (granted in 2012). She has published several papers in JCR journals and international conference communications. She has participated in several research projects (with public and private fundings) related to healthcare data-driven machine learning systems. Her current research interests include machine learning, data science, and statistical learning theory. She won the Orange Foundation Best Ph.D. Thesis Award by the Spanish Official College of Telecommunication Engineering.

A. P. ENGELBRECHT (Senior Member, IEEE) received the master's and Ph.D. degrees in computer science from Stellenbosch University, South Africa, in 1994 and 1999, respectively. Prior to 2019, he was with the Department of Computer Science, University of Pretoria, from 1998 to 2018, where he served as the Head of the Department, from 2008 to 2017; the South African Research Chair of artificial intelligence, from 2007 to 2018; and the Director of the Institute for Big Data and Data Science, from 2017 to 2018. He is currently appointed as the Voigt Chair of data science with the Department of Industrial Engineering, with a joint appointment as a Professor with the Computer Science Division, Stellenbosch University. He is the author of two books *Computational Intelligence: An Introduction* and *Fundamentals of Computational Swarm Intelligence*. He is the author/co-author of over 370 articles. He currently holds an NRF A-rating. His research interests include swarm intelligence, evolutionary computation, artificial neural networks, artificial immune systems, machine learning, data analytics, and the application of these artificial intelligence paradigms to data mining, games, bioinformatics, finance, and difficult optimization problems.

PABLO DE MIGUEL BOHOYO received the degree in business management and administration from the Complutense University of Madrid, Spain, in 1998, and the Ph.D. degree in health economy in conjunction with Rey Juan Carlos University, in 2012. He is currently the Control Management Manager with the University Hospital of Fuenlabrada. He has participated in several research projects (with public and private fundings) related to healthcare economy. His current research interests include data science and equity and efficiency in the field of health economy.

INMACULADA MORA-JIMÉNEZ received the degree in telecommunication engineering from the Polytechnic University of Valencia, Spain, in 1998, and the Ph.D. degree in telecommunication from the University Carlos III of Madrid, Spain, in 2004. She is currently an Associate Professor with the Department of Signal Theory and Communications, Telematics and Computing Systems, Rey Juan Carlos University, Spain. She has conducted her research mainly in data analytics and biomedical engineering. She is a coauthor of 40 JCR indexed articles and more than 50 contributions to international conferences. She has participated in 18 competitive research projects (principal investigator of four) and collaborated in more than 20 projects with private funding entities. Her main research interests include data science and machine learning with application to image processing, bioengineering, and wireless communications.