Journal of Medical Systems (2021) 45: 31
https://doi.org/10.1007/s10916-020-01704-5

MOBILE & WIRELESS HEALTH

A Hybrid Model to Classify Patients with Chronic Obstructive Respiratory Diseases

Diogo Martinho[1] · Alberto Freitas[2] · Ana Sá-Sousa[2] · Ana Vieira[1] · Jorge Meira[1] · Constantino Martins[1] · Goreti Marreiros[1]

Received: 9 October 2020 / Accepted: 27 December 2020 / Published online: 30 January 2021

Abstract

Over the last decades, an increase in the ageing population and age-related diseases has been observed, with the increase in healthcare costs. As so, new solutions to provide more efficient and affordable support to this group of patients are needed. Such solutions should never discard the user and instead should focus on promoting more healthy lifestyles and provide tools for patients' active participation in the treatment and management of their diseases. In this concern, the Personal Health Empowerment (PHE) project presented in this paper aims to empower patients to monitor and improve their health, using personal data and technology assisted coaching. The work described in this paper focuses on defining an approach for user modelling on patients with chronic obstructive respiratory diseases using a hybrid modelling approach to identify different groups of users. A classification model with 90.4% prediction accuracy was generated combining agglomerative hierarchical clustering and decision tree classification techniques. Furthermore, this model identified 5 clusters which describe characteristics of 5 different types of users according to 7 generated rules. With the modelling approach defined in this study, a personalized coaching solution will be built considering patients with different necessities and capabilities and adapting the support provided, enabling the recognition of early signs of exacerbations and objective self-monitoring and treatment of the disease. The novel factor of this approach resides in the possibility to integrate personalized coaching technologies adapted to each kind of user within a smartphone-based application resulting in a reliable and affordable alternative for patients to manage their disease.

Keywords User Modelling · Personalized coaching · Mobile health · Preventive healthcare · Healthcare management systems

Introduction

Rising costs of healthcare due to the ageing population and related increase of non-communicative diseases urges for finding ways to save expenses by diminishing the need for care and making the current care more efficient [1]. At present, healthcare provision is reactive and process driven, and patients are treated according to predefined pathways and hardly consider individual necessities and capabilities [2]. As a result, health authorities and healthcare providers are noticing the patient or person

This article is part of the Topical Collection on *Mobile & Wireless Health*

✉ Diogo Martinho
diepm@isep.ipp.pt

Alberto Freitas
alberto@med.up.pt

Ana Sá-Sousa
anasasousa@gmail.com

Ana Vieira
aavir@isep.ipp.pt

Jorge Meira
janme@isep.ipp.pt

Constantino Martins
acm@isep.ipp.pt

Goreti Marreiros
mgt@isep.ipp.pt

[1] Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development (GECAD), Institute of Engineering, Polytechnic of Porto, Porto, Portugal

[2] CINTESIS – Center for Health Technology and Services Research, Porto, Portugal

resource that had remained unused until now. By starting with the primary need of the person – to be healthy – and including him/her into the process in an active role, new paradigms for healthcare become possible. Significant cost reductions can be achieved by developing preventive solutions to help the person adopt a healthier lifestyle – thus reducing the number of patients – and by providing the person with tools to actively participate in the treatment when diseases do arise – thus decreasing the burden on care personnel [3]. In this context, we have recently observed great advances in technology to empower people self-care [4]. Smartphones and tablets and quantified self-style self-monitoring wellness devices are commonplace more than ever before. Wellness oriented solutions often suffer from short-term use due to quickly diminishing interest from their users [5] and from lack of possibilities to utilize them in conjunction with clinical healthcare treatments [6]. Patients are left alone with their problems in between therapy or treatment, and the possibly collected personal data is left unused. Insights in psychology and gamification allow for approaches to keep people engaged and achieve the desired health impact [7].

The present work is enclosed in the Personal Health Empowerment project (PHE) (https://itea3.org/project/personal-health-empowerment.html). The PHE has the main goal to empower people with Chronic Obstructive Respiratory Diseases (CORD) and to help patients monitor and improve their health using personal data and technology assisted coaching (also known as digital coaching). Some of the innovations in the PHE project include the definition and development of reliable ways to self-measure pain, respiratory function, and behaviour development of analytics on heterogeneous personal health sources. All these variables will provide insight in the relation between behaviour and health specification of methodologies. As a result, interactive, dynamic, and personalized coaching programmes will be developed with the definition of innovative and motivating approaches for long-term adherence, bridging the gap between wellness and care. In this work we present an overview on existing approaches for user modelling according to literature which includes studying available characteristics and techniques to model different user profiles. Then, and as the main goal of this work, the most relevant characteristics and techniques will be identified and integrated into a user modelling approach for patients with CORD, also including information such as: general health data, habits, behaviours, symptoms, diagnosis, historical treatments and therapies of the patient. The proposed user model will be essential to establish the viability and constraints to develop a coaching platform to provide efficient recommendations to patients with CORD with different necessities and capabilities.

In the next section, literature on user modelling is reviewed. In Section 3, it is described the clustering analysis performed for the CORD Management use case and finally, major conclusions are taken in Section 4 as well as the work to be done hereafter.

User Modelling

This section describes different definitions associated with user modelling which includes the main characteristics and techniques existing in the literature.

A user model is composed by a set of characteristics that adjust the content, presentation and navigation to each user. These characteristics can be domain-dependent and domain-independent and are related to beliefs, preferences, knowledge and attributes about the user. Domain dependent (DD) data is related with system responses tailored according to the domain knowledge of a user [8]. For this, it is necessary to perceive user current state and knowledge regarding concepts and relations inherent to the domain, predict how the user will interpret system responses, understand the many different goals and plans of each user, predict and respond to different mistakes while the user is using the system and identify the most adequate way to present information to each user. Different methods can be used to measure user knowledge and expertise regarding the domain: Direct Dialogue and Indirect Acquisition. Direct Dialogue is performed directly with the user in order to assess his/her expertise in the domain. Direct Dialogue features allow users to input and share their knowledge (for example, using questionnaires or forms) and include mechanisms to process the inserted data and measure user knowledge regarding the domain. Indirect acquisition method allows the system to assess user knowledge indirectly according to how the user performs different actions. Depending on this assessment the user knowledge regarding the domain is classified on different levels which in turn are updated over time as the user works with the system. Domain independent (DI) data is not related to user expertise regarding the domain but to his/her cognitive abilities which indicate how the user perceives, thinks, remembers, behaves and solves different problems [8]. In other words, domain-independent knowledge corresponds to the phycological characteristics of the user. There are many different psychological models and tests that can be used to assess user personality such as the Myer-Briggs Type Indicator [9], the Eysenck's Pen Model [10] and the Big Five Model [11–14].

After identifying the data related to each user's characteristics, it is then possible to define the algorithms that will process this data and in turn affect the computational environment. These algorithms are mainly defined using statistical and non-statistical techniques. Among different statistical techniques it is highlighted the use of Beta Distribution [15], Linear Modelling, Markov Model, Bayesian Networks and Rule Induction with statistical data [16]. Examples of non-statistical techniques include the use of an Overlay Model [17, 18], Perturbation Model [17, 19], Knowledge Modelling, Behaviour-based Model [20], Rule-based Model, Stereotypes [21] and Ontologies [22–26].

Personal health empowerment

Two of the greatest challenges associated with traditional healthcare systems are related to the fact that they are reactive and process driven [2].This means that these systems often treat patients according to predefined pathways with limited possibilities to consider individual necessities and capabilities. Furthermore, healthcare tends to be provided to people only when they are diagnosed with certain health issues (reactive), instead of assessing risks in time and preventing adverse health development (proactive).

PHE project is addressing these challenges in healthcare by focusing on patient support and allowing him/her to better manage his/her own health-related behaviour and participate actively in the treatment of the disease (https://itea3.org/project/personal-health-empowerment.html). Wellness oriented solutions usually do not last very long as patients quickly lose interest to keep using these solutions in conjunction with clinical healthcare treatments. Therefore, in the current paradigm, patients are left alone with their problems in between therapy or treatment sessions and any personal data collected is left unused. The main goal of PHE is to change the current observed paradigm by empowering people to monitor and improve their health using personal data and technology assisted coaching. For this, PHE will help to provide both evidence and means to realize people-centric and preventive healthcare and allow for cost-saving self- and home-care solutions with increased patient involvement.

The project intends to apply the innovative measurement, monitoring, heterogeneous data analysis and intelligent coaching solutions to different use cases: Healthy Workplaces and CORD Management. The first use case is related to the necessity to provide ideal workplace conditions from a wellbeing point of view to bring improvements to both the physical and mental health of workers. The second use case is related to CORD which are a public health problem with increasing demands on healthcare systems. Nowadays, there is a growing market demand for solutions that can help to reduce costs, while maintaining the quality of care [27, 28]. Patients with CORD are continuously at risk of deterioration of health, requiring regular medical check-ups and monitoring of their health status. Traditionally health care is delivered through clinicians' face-to-face interaction. With the growing prevalence of CORD and continuous pressure from healthcare authorities/insurance companies, an increasing number of patients are being managed at home in their own environment and most of the time being left alone with traditional self-management materials (books, leaflets, videos, and web-based technology) [29–31]. To overcome the limitations of traditional healthcare systems, new solutions are now being developed such as the development of coaching solutions and mHealth technologies. Coaching solutions appear to be an ideal platform to deliver both simple and effective self-management interventions while maintaining/improving the quality of care and reducing costs [32]. mHealth technologies for CORD should involve monitoring and managing signs and symptoms of the disease, empowering patients to recognize the early signs of exacerbations and to develop skills to better manage their disease. Several monitoring systems have been proposed in the context of CORD management over the last years, but these show evident limitations that should be discussed. A great number of existing proposals already combine different machine learning techniques in order to monitor the health condition of the patient [33–38] and provide personalized interactions. In this sense, we have seen systems using techniques such as fuzzy classifiers, artificial neural networks [34, 37–39], reinforcement learning [40, 41], among others. However, in the context of CORD management, this monitorization is mainly performed with the goal to analyse patient data and detect respiratory diseases or respiratory complications such as exacerbations [36–38] rather than understanding the profile (and associated behaviours) of the patient and anticipating further complications. This means that in some way most existing systems are less proactive and more reactive to the current health condition of the patient. Another issue which may compromise the usability of this kind of systems is that it often requires the use of multiple devices such as a smartphone solution combined with a digital spirometer to analyse respiratory function [36–38], which bring additional costs to the average user. With these points in mind, PHE project was designed to provide a solution which relies only upon the use of a smartphone and its embedded sensors to correctly monitor and capture patient data through the application of innovative monitoring algorithms. Furthermore, in this work, we present how patient data captured through the use and interaction with the smartphone will be analysed using machine learning (as will be explained in the following section) to identify different user profiles. This analysis is performed not to detect whether a patient has a certain respiratory disease but to identify through his/her health behaviours the type of user being monitored and then be able to support him more efficiently.

Description of data

In this section, it is presented an overview of all the user characteristics that have been considered for the CORD Management use case. Each identified characteristic is related to either DD or DI data and can be retrieved through different tools such as questionnaires and self-reported data (user input), healthcare records, clustering analysis, etc. Table 1 shows all relevant user characteristics identified for the CORD Management use case. Descriptions and examples of

Table 1 User Characteristics for the CORD Management Divided in Domain-Dependent and Domain-Independent Data

	Characteristics	Descriptions/Examples	Tools to Collect Data
DI Data	Personal information	Name, Email, Password	User input
	Demographic data	Age, BMI, Sex	User input
	Patient background	Smoking habits, Pregnancy, allergen sensitisation	User input
	Diagnosis	Respiratory disease (asthma, COPD) and comorbidities	User input, Healthcare records
	Domain of application	Geographic localization of the user	Smartphone sensors (GPS)
	Knowledge (background knowledge)	A collection of knowledge translated in concepts. Possibility of a qualitative, quantitative or probabilistic indication of concepts and knowledge acquired for the user	User input
	Cognitive capabilities	Emotional state (anxiety, depression, stress, etc.,)	Psychological exams, User input
DD Data	Objectives	User objectives regarding the use of the system	User input
	Personal preferences	Classifications of recommendations provided (useful, not useful), Interests (hobbies, routines)	User input
	Complete description of the navigation	Kept register of each page accessed, capacity to use the system, definition of the individual preferences with the objectives to adapt the navigation and contents	User input, Adaptative interfaces
	Knowledge acquired	A collection of knowledge translated in concepts.	Expert input
	Medication use and Health status	Data related to patient intake of medication; inhalations; record symptoms and exacerbations (rescue medication, hospitalization)	User input, Computerised Respiratory Auscultation, Healthcare records
	Context model	Data related with the environment of the user (localization of the user); Existence of caregiver or isolated user	External resources (Public APIs)
	Activity tracking	Kept register of end users' daily activity	External resources (Smartphones, Google Fit platform)

BMI: Body Mass Index; COPD: Chronic Obstructive Pulmonary Disease.

each identified characteristic are also provided as well as the tools will be used to collect that information.

All the information related to each characteristic also includes the specification of different high-level and low-level variables that characterize the user's current health condition and his/her surrounding environment. Examples of these variables are: 1) Body Mass Index, which corresponds to the value derived from the mass (weight) and height of an individual; 2) Use of Continuous Positive Airway Pressure, which is a form of positive airway pressure ventilator that continuously applies mild air pressure to keep the airways open in people who are not able to breathe spontaneously on their own; 3) Obstructive Sleep Apnea, which is caused by complete or partial obstructions of the upper airway during sleep; 4) Air Quality Index, which indicates how polluted the air currently is or how polluted it is forecast to become; etc.

Clustering analysis

A hybrid model to perform the Clustering Analysis for the CORD Management use case was defined by combining two different clustering models (hierarchical and classification models). The architecture of the proposed hybrid model is presented in Fig. 1. It comprises a process with four main steps, which will be explained in more detail in the following sections.

Data description

The proposed hybrid model will be validated using data obtained from the Control and Burden of Asthma and Rhinitis (ICAR) study (PTDC/SAU-SAP/119192/2010), a nation-wide population-based observational cross-sectional study conducted in Portugal (ClinicalTrials.gov: NCT01771120) [42]. Included participants (n=726) were from the general population and aged 18 years and older. The mean age of the participants was 44 years old, and 63% (n=469) of the participants were females. For each patient, it was collected data on lung function and exhaled nitric oxide, skin prick tests, a structured clinical assessment, and standardized questionnaires. The data collected comprised a total of 1181 variables described in different data formats (numeric, classification, binary, etc.).

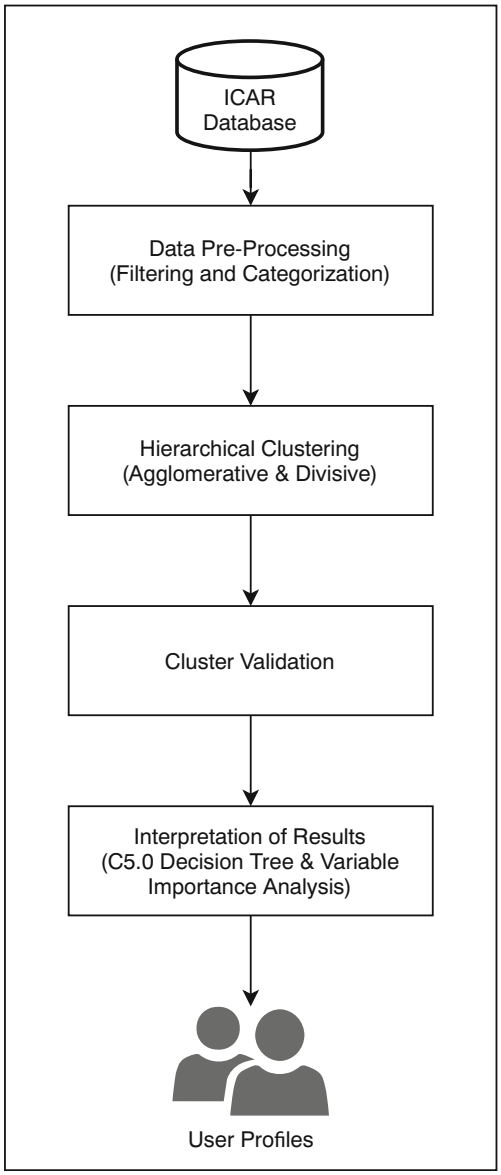

Fig. 1 Architecture of the Proposed Hybrid Model Divided in 4 Main Steps (Data Pre-Processing, Hierarchical Clustering, Cluster Validation, and Interpretation of Results – Classification Model)

Data pre-processing

The first step of the clustering analysis performed in this study was the pre-processing of the data from the ICAR study. To fit the data into the context of the CORD Management use case, the pre-processing activities focused on two steps: data filtering and categorization. In the first step, a manual data filtering was performed to exclude noisy data, i.e., the variables of the ICAR study that were not considered for the CORD Management use case were excluded. In one hand, the total number of variables that comprise CORD Management is of 253 variables and whose description has been provided in Table 1. On the other hand, and as mentioned above, ICAR study collected patient data on 1181 variables. The analysis performed in this study consisted of accessing (1) patient data on variables that are also considered in CORD Management use case, which comprise a total of 96 independent variables, and among those variables (2) exclude any shown with single values. Finally, 93 variables were considered for further analysis. After identifying these variables, the second step of Data Pre-Processing was performed by categorizing each variable according to established categories within medical literature. For instance, GA2LEN Score [42], has established medical scores which describe the probability of a patient having Asthma (a score of 0 suggests the patient is unlikely to have asthma, a score between 1 and 3 suggest the patient may have asthma and further diagnosis is required, and a score of 4 or higher suggests the patient is very likely to have asthma).

Hierarchical clustering

After pre-processing, hierarchical clustering was applied to the dataset to identify different clusters. Firstly, it was necessary to define a dissimilarity matrix in which the distance between different points – i.e. different patients - is measured using a distance function. For this study, the dissimilarity matrix was defined using Gower Distance [43] which is capable of handling categorical data and measuring the distance between two instances X_i and X_j differently according to each considered variable. For that, the following formula is applied:

$$s_{ij} = \frac{\sum_{k=1}^{N} w_{ijk} s_{ijk}}{\sum_{k=1}^{N} w_{ijk}}$$

Where w_{ijk} corresponds to the weight for a variable k between the instances X_i and X_j, and s_{ijk} corresponds to the difference between the value of a variable k for both instances X_i and X_j.

After defining the dissimilarity matrix, hierarchical clustering can then be performed. The two most common hierarchical clustering algorithms are the agglomerative and divisive clustering. Agglomerative clustering (AC) is a bottom-up approach in which each instance is firstly treated as a single cluster and then pairs of clusters are merged successively until one cluster containing all instances is defined. Divisive clustering (DC) is a top-down approach in which a cluster is firstly defined containing all instances and then the most heterogenous cluster is iteratively divided until each instance is a cluster. The resulting dendrograms containing all clusters formed using both agglomerative and divisive approaches are presented in Figs. 2 and 3, respectively.

Observing the two generated dendrograms, the identification of the algorithm that better suits the data was not clear, and an additional step was required. The cluster validation and the identification of the ideal number of clusters and the most

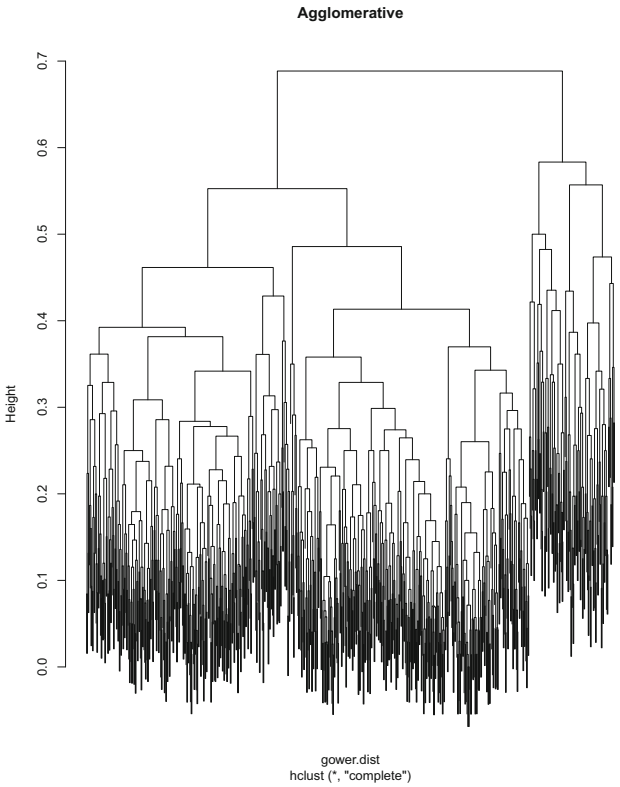

Fig. 2 Dendrogram generated using Agglomerative Clustering for the CORD Management Use Case

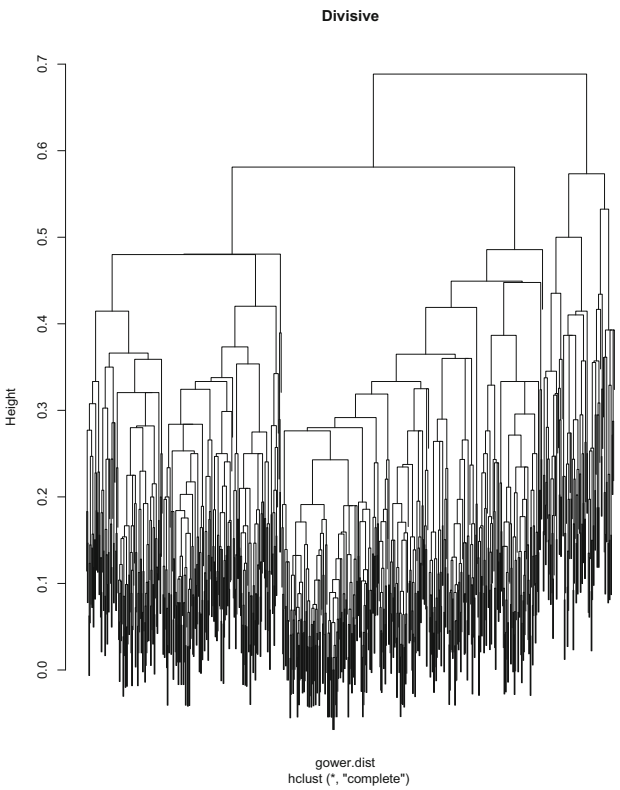

Fig. 3 Dendrogram generated using Divisive Clustering for the CORD Management Use Case

suitable clustering algorithm are explained in the following section.

Cluster validation

In the Clustering Validation phase, each algorithm and generated clusters were validated in terms of size and distance between each cluster, based on the within sum of squares value. This sum serves as a measure to express how close the instances are within a cluster. In other words, the lower the within sum of squares value is, the closer the instances are within the clusters. To visualize this distribution, an elbow curve graph was created showing the within sum of squares regarding the generated clusters for each clustering algorithm. These graphs are shown in both Figs. 4 and 5.

In case of AC, a bend is observed at 5 clusters and after that there is a significant decrease of the within sum of squares value at 10 clusters. Regarding DC, a bend is observed at 3 clusters and after that there is also a significant decrease of the within sum of squares value at 10 clusters. To have a better idea of what a good number of clusters is for this algorithm and which Clustering method should be considered, the generated clusters should be also analysed in terms of size.

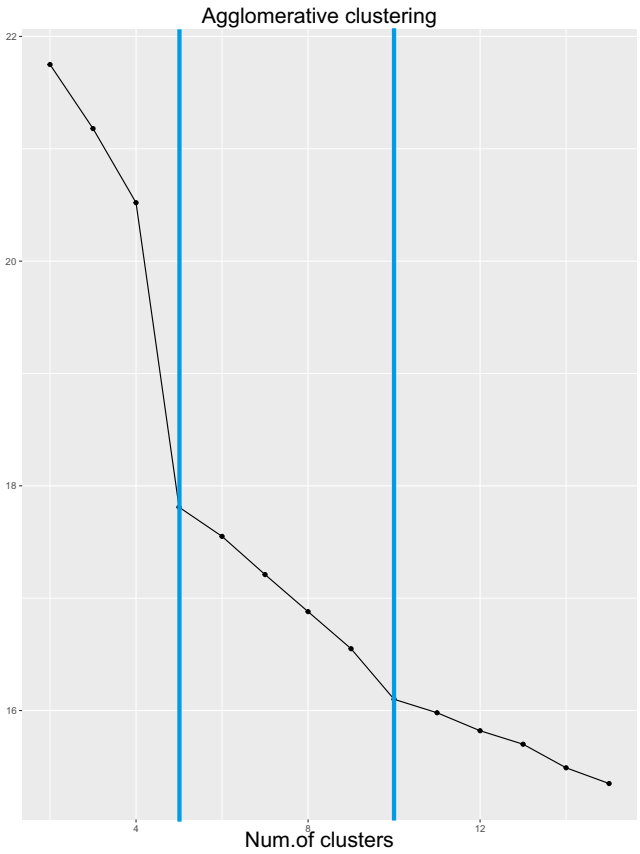

Fig. 4 Elbow Curve for Agglomerative Clustering (Bends Observed at 5 and 10 Clusters)

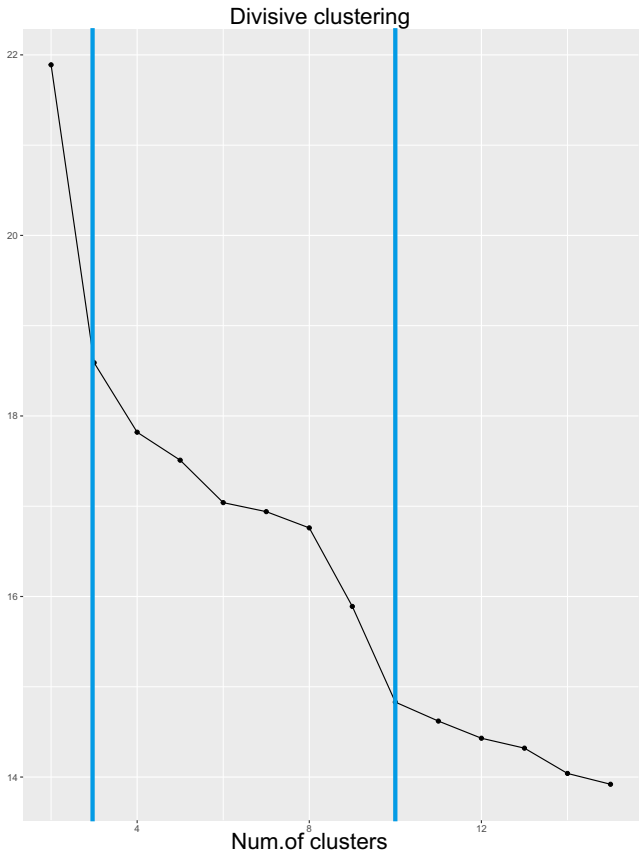

Fig. 5 Elbow Curve for Divisive Clustering (Bends Observed at 3 and 10 Clusters)

Looking at Figs. 6 and 7, the observations seem similarly balanced for both approaches. With a small cluster number, both approaches are fairly distributed. DC shows a gap of 360 observations in Cluster-2 to 99 observations in Cluster-3 while AC shows a gap of 329 observations in Cluster-2 to 28 observations in Cluster. If the number of clusters increases, DC shows two clusters with very few observations (Cluster-9 with 4 observations and Cluster-10 with only 1 observation) while AC shows clusters with only a few more observations (the smallest cluster is Cluster-7 with 9 observations).Imbalanced clusters could lead to more biased comparisons as some clusters with more instances could outweigh the remaining clusters [44]. For this reason, both approaches seem more adequate with smaller cluster sizes. Furthermore, since there is no clear difference between both approaches with smaller sizes (3 clusters for DC and 5 clusters for AC) it was chosen 5 clusters using AC for the remainder of this analysis given the large size of independent variables considered after the data pre-processing phase.

Classification model

After identifying and generating the ideal number of clusters, these were classified using C5.0 Decision-Tree algorithm. This algorithm allows to identify, among the 93 independent variables included for this study, those which had more weight in the division of each instance of the defined clusters. In this sense, a cross-validation was necessary to avoid either over and under-fitting the decision model, and this process consisted of training and validating the filtered data from ICAR study over 10 iterations. For each iteration, 66% of the total amount of data was used to train the decision model while 33% was used to validate the decision model. In this 10-fold cross-validation analysis, the classification model with highest accuracy was selected.

According to Fig. 8, the classification model with the highest accuracy (91.4%) was the model obtained in the first iteration. The remaining iterations with higher accuracy levels were iteration 3 (87.3% accuracy) and iteration 10 (89.4%).

As can be seen in Table 2 iteration 1 cluster division considered 7 of the 93 independent variables and 7 classification rules (Table 3) that describe the characteristics of each cluster. Regarding variables usage among iterations with higher accuracy (iteration 1, 3 and 10), the variables most used were “Number of inhalers”, “Sensitization to at least one indoor allergen” and “At least one factor for asthma related death”, which were all used in the cluster division for each one of these iterations. After that, the two variables most used for cluster division were “Number of asthma exacerbations on the previous year” (iteration 3 and 10), and “Bronchodilator reversibility based on previous spirometry” (iteration 1 and 10). In fact, all variables considered in iteration 10 were also considered in either iteration 1 or iteration 3. Furthermore, iteration 3 had the highest

Fig. 6 Cluster Size for Divisive Clustering (Size of Each Cluster Highlighted for 3 and 10 Clusters Division)

Cluster- 1	size	629.00	269.00	269.00	269.00	269.00	269.00	265.00	105.00	105.00	105.00	105.00	105.00
Cluster- 2	size	99.00	360.00	360.00	360.00	360.00	359.00	359.00	359.00	270.00	270.00	270.00	270.00
Cluster- 3	size	0.00	99.00	62.00	62.00	27.00	27.00	27.00	160.00	89.00	85.00	85.00	85.00
Cluster- 4	size	0.00	0.00	37.00	21.00	35.00	35.00	35.00	27.00	160.00	160.00	160.00	149.00
Cluster- 5	size	0.00	0.00	0.00	16.00	21.00	21.00	21.00	35.00	27.00	27.00	21.00	21.00
Cluster- 6	size	0.00	0.00	0.00	0.00	16.00	16.00	16.00	21.00	35.00	35.00	35.00	35.00
Cluster- 7	size	0.00	0.00	0.00	0.00	0.00	1.00	4.00	16.00	21.00	4.00	4.00	4.00
Cluster- 8	size	0.00	0.00	0.00	0.00	0.00	0.00	1.00	4.00	16.00	21.00	21.00	18.00
Cluster- 9	size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	4.00	16.00	16.00	16.00
Cluster- 10	size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	4.00	6.00	6.00
Cluster- 11	size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	4.00	3.00
Cluster- 12	size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	4.00
Cluster- 13	size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
Cluster- 14	size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Cluster- 1	size	610.00	610.00	610.00	281.00	281.00	281.00	281.00	281.00	231.00	231.00	231.00	231.00	231.00
Cluster- 2	size	118.00	49.00	49.00	329.00	329.00	319.00	319.00	319.00	50.00	50.00	50.00	50.00	12.00
Cluster- 3	size	0.00	69.00	28.00	49.00	39.00	39.00	28.00	28.00	319.00	319.00	319.00	319.00	319.00
Cluster- 4	size	0.00	0.00	41.00	28.00	28.00	28.00	28.00	28.00	28.00	28.00	7.00	7.00	7.00
Cluster- 5	size	0.00	0.00	0.00	41.00	41.00	41.00	41.00	32.00	28.00	28.00	21.00	21.00	21.00
Cluster- 6	size	0.00	0.00	0.00	0.00	10.00	10.00	11.00	9.00	32.00	32.00	28.00	24.00	24.00
Cluster- 7	size	0.00	0.00	0.00	0.00	0.00	10.00	10.00	11.00	9.00	4.00	32.00	32.00	38.00
Cluster- 8	size	0.00	0.00	0.00	0.00	0.00	0.00	10.00	10.00	11.00	11.00	4.00	4.00	32.00
Cluster- 9	size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.00	10.00	10.00	11.00	11.00	4.00
Cluster- 10	size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.00	10.00	10.00	4.00	11.00
Cluster- 11	size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.00	10.00	10.00	4.00
Cluster- 12	size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.00	10.00	10.00
Cluster- 13	size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.00	10.00
Cluster- 14	size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.00

Fig. 7 Cluster Size for Agglomerative Clustering (Size of Each Cluster Highlighted for 5 and 10 Clusters Division)

number of variables considered in cluster division (13 of the 93 Independent variables).

Results obtained in the first iteration and after generating the decision tree reveal the importance of the variables in the division of each instance per cluster. The variable with the highest usage percentage was the use (or not) of inhalers. Within the group of patients who used inhalers (clusters 3 to 5), the most relevant variable was the presence (or not) of at least one risk factor for asthma-related death. Cluster 3 is characterized by having at least one risk factor for asthma-related death including severe asthma exacerbation in the last year and cluster 5 is characterized by having at least one risk factor for asthma-related death but not severe asthma exacerbation in the last year and clusters 4 by none of these two variables. The group of patients who did not use of inhalers is further characterized by the sensitization to at least one indoor allergen (cluster 1) and by the bronchodilator reversibility based on previous spirometry (cluster 2).

Clusters distribution seems to suggest that cluster 1 groups patients with allergies but not any respiratory disease; cluster 2 groups patients that although probably having asthma or COPD do not use inhalers to control their disease; cluster 3 groups patients with moderate to severe rhinitis and asthma that is not controlled despite the use of inhalers; cluster 4 groups patients with the use of inhalers for respiratory disease, and without exacerbations; and cluster 5 groups patients with mild rhinitis and asthma that is not controlled despite the use of inhalers.

Each obtained cluster will allow us to define specific user profiles within the scope of the CORD Management use case which in turn can provide adjusted content to the patient based on his profile. This includes personalized recommendations and coaching plans targeted at specific user profiles.

Conclusions and future work

CORD are a public health problem with increasing demands on healthcare systems. Nowadays, there is a growing market demand for solutions which can help to reduce costs, while maintaining the quality of care. Patients with CORD are continuously at risk of deterioration of health, requiring regular medical check-ups and monitoring of their health status. Traditionally health care is delivered through clinicians' face-to-face interaction. With the growing prevalence of CORD and continuous pressure from healthcare authorities/insurance companies, an increasing number of patients is being managed at home in their own environment and most of the time being left alone with traditional self-management materials (books, leaflets, videos, and web-based technology). To

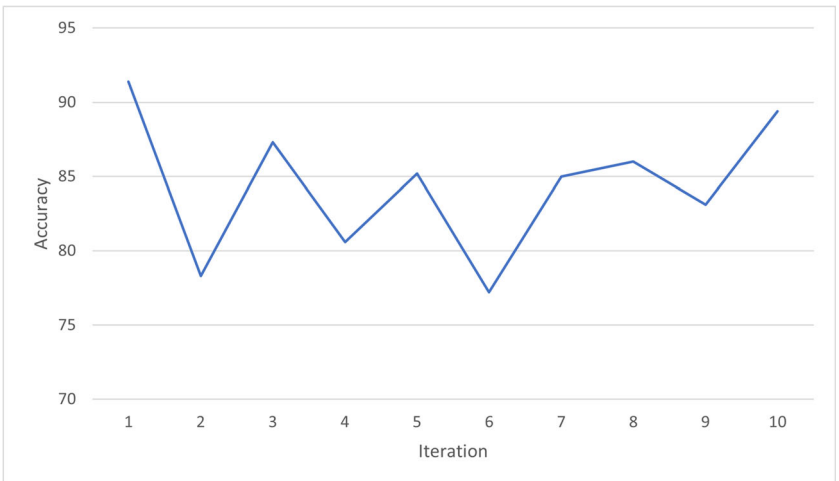

Fig. 8 Cross-validation Analysis (Highest Accuracy Values Correspond to Iterations 1, 3 and 10)

Table 2 Independent Variables Usage for Iteration 1, 3 and 10

Characteristics	Iteration (s)	Independent Variable	Code and Value
Demographic Data	3	Age	1: > 0; 2: > 20; 3: > 40; 4: > 60; 5: > 80
	3	BMI	1: <18.5; 2: < 24.99; 3: >=25; 4: >=30
Patient background	**1**, 3, 10	Sensitization to at least one indoor allergen*	0: No; 1: Yes
	1	Severe asthma exacerbation in the last 12 months**	0: 0; 1: >= 1
	1, 3, 10	At least one risk factor for asthma related death***	0: No; 1: Yes
	3, 10	Number of asthma exacerbations on the previous year	0: 0; 1: <11; 2: <21; 3: <51; 4: >50
	3	Follow-up by a specialist	0: No; 1: Yes
	3	Impairment in leisure activities due to rhinitis	0: No; 1: Yes
	3	Presence of major psychological problems	0: No; 1: Yes
Diagnosis	3	Asthma	0: No; 1: Yes
	3	Asthma screening (GA2LEN Score)	0: 0; 1: < 4; 2: >=4
	1	Rhinitis severity	1: Mild; 2: Moderate-severe
	1, 3	Bronchodilator reversibility based on previous spirometry	0: No; 1: Yes
Medication use	**1**, 3, 10	Number of inhalers	0: 0; 1: >= 1
	1	Number of rescue medication for respiratory disease	0: 0; 1: >= 1
	3	Number of ICS	0: 0; 1: >= 1

* at least one positive: mite allergy, animal epithelia allergy, mold allergy. **hospitalization for asthma in the past year; emergency care visit for asthma in the past year. *** presence of at least one: history of intensive care unit admission, mechanical ventilation, anaphylaxis and/or confirmed food allergy; hospitalization for asthma in the past year; emergency care visit for asthma in the past year; use oral corticosteroids or over-use of SABAs without use of inhaled corticosteroids; poor adherence with asthma medications or lack of a written asthma action plan. BMI: Body Mass Index; ICS: Inhaled Corticosteroids

overcome the limitations of traditional healthcare systems, new solutions are now being developed such as the development of coaching solutions and mHealth technologies. Coaching solutions appear to be an ideal platform to de-liver both simple and effective self-management interventions, while maintaining/improving quality of care and reducing costs. mHealth technologies for CORD should involve monitoring and managing signs and symptoms of the dis-ease, empowering patients to recognize the early signs of exacerbations and to develop skills to be active participants in management and treatment of their disease.

Table 3 Classification Rules Generated for Iteration 1

Rule	Cluster	Condition(s)
1	1	Number of inhalers = 0
		Sensitization to at least one indoor allergen = 1
2	2	Number of inhalers = 0
		Bronchodilator reversibility based on previous spirometry = 1
3	2	Number of inhalers = 0
4	3	At least one risk factor for asthma related death = 1
		Rhinitis severity = 2
		Number of rescue medication for respiratory disease = 1
5	3	At least one risk factor for asthma related death = 1
		Number of Inhalers = 1
		Severe asthma exacerbation in the last 12 months = 1
6	4	At least one risk factor for asthma related death = 0
		Number of Inhalers = 1
7	5	At least one risk factor for asthma related death = 1
		Number of inhalers = 1
		Severe asthma exacerbation in the last 12 months = 0

In this paper it was presented an overview on user modelling including characteristics and techniques most frequently used to model different users. The PHE project was presented and the clustering analysis to identify different clusters of users for the CORD Management use case was defined. For this, several required steps were described, and it was possible to identify 7 different rules which describe a total of 5 clusters of users within the CORD Management use case. The obtained clusters will be essential in the development of the PHE healthcare solution in terms of building a more personalized coaching solution that can adapt all provided recommendations according to the characteristics of each user which ultimately make it possible to enhance and improve user current health condition and promote a healthier lifestyle.

As future work we intend to validate and test the personalized solutions which will be developed starting with single-visit validation in a clinical environment and finally with observational validation in a community environment. Ultimately, the goal will be to assess the potential of self-monitoring and personalization on reducing exacerbations and symptoms when used in between healthcare appointments. We believe that data generated during validation period will also create new opportunities for research and for the development of further innovative healthcare services to better support patients suffering from chronic obstructive respiratory diseases.

Acknowledgments The work presented in this paper has been developed under the EUREKA - ITEA3 Project PHE (PHE-16040), and by National Funds through FCT (Fundação para a Ciência e a Tecnologia) under the project UIDB/00760/2020 and by NORTE-01-0247-FEDER-033275 (AIRDOC - "Aplicação móvel Inteligente para suporte individualizado e monitorização da função e sons Respiratórios de Doentes Obstrutivos Crónicos") by NORTE 2020 (Programa Operacional Regional do Norte).

References

1. Broens, T., Van Halteren, A., Van Sinderen, M., Wac, K.: Towards an application framework for context-aware m-health applications. International Journal of Internet Protocol Technology 2, 109-116 (2007)
2. Hii, P.-C., Chung, W.-Y.: A comprehensive ubiquitous healthcare solution on an Android™ mobile device. Sensors 11, 6799-6815 (2011)
3. Agnihothri, S., Cui, L., Delasay, M., Rajan, B.: The value of mHealth for managing chronic conditions. Health care management Science 23, 185-202 (2020)
4. Iftikhar, S., Ahmad, F., Fatima, K.: A Semantic Methodology for Customized Healthcare Information Provision. Information Sciences Letters 1, 49-59 (2012)
5. Woldaregay, A.Z., Issom, D.-Z., Henriksen, A., Marttila, H., Mikalsen, M., Pfuhl, G., Sato, K., Lovis, C., Hartvigsen, G.: Motivational Factors for User Engagement with mHealth Apps. In: pHealth, pp. 151-157. (2018)
6. Sobnath, D.D., Philip, N., Kayyali, R., Nabhani-Gebara, S., Pierscionek, B., Vaes, A.W., Spruit, M.A., Kaimakamis, E.: Features of a mobile support app for patients with chronic obstructive pulmonary disease: literature review and current applications. JMIR mHealth and uHealth 5, e17 (2017)
7. Johnson, D., Deterding, S., Kuhn, K.-A., Staneva, A., Stoyanov, S., Hides, L.: Gamification for health and wellbeing: A systematic review of the literature. Internet interventions 6, 89-106 (2016)
8. Myers, I.B.: The Myers-Briggs Type Indicator: Manual (1962). (1962)
9. Eysenck, H.J.: Dimensions of personality. Transaction Publishers (1950)
10. Fiske, D.W.: Consistency of the factorial structures of personality ratings from different sources. The Journal of Abnormal and Social Psychology 44, 329 (1949)
11. Norman, W.T.: Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. The Journal of Abnormal and Social Psychology 66, 574 (1963)
12. McCrae, R.R., John, O.P.: An introduction to the five-factor model and its applications. Journal of personality 60, 175-215 (1992)
13. Goldberg, L.R.: An alternative" description of personality": the big-five factor structure. Journal of personality and social psychology 59, 1216 (1990)
14. Orwant, J.: For want of a bit the user was lost: Cheap user modeling. IBM Systems Journal 35, 398-416 (1996)
15. Zukerman, I., Albrecht, D.W.: Predictive statistical models for user modeling. User Modeling and User-Adapted Interaction 11, 5-18 (2001)
16. Kass, R., Finin, T.: Modeling the user in natural language systems. Computational Linguistics 14, 5-22 (1988)
17. Martins, C., Faria, L., De Carvalho, C.V., Carrapatoso, E.: User modeling in adaptive hypermedia educational systems. Educational Technology & Society 11, 194-207 (2008)
18. Nguyen, L., Do, P.: Combination of Bayesian network and overlay model in user modeling. International Conference on Computational Science 5-14 (2009)
19. Bushey, R., Mauney, J.M., Deelman, T.: The development of behavior-based user models for a computer system. UM99 User Modeling, pp. 109-118. Springer (1999)
20. Rich, E.: User modeling via stereotypes. Cognitive science 3, 329-354 (1979)
21. Brickley, D.: RDF vocabulary description language 1.0: RDF schema. http://www.w3.org/TR/rdf-schema/ (2004)
22. McGuinness, D.L., Van Harmelen, F.: OWL web ontology language overview. W3C recommendation 10, 2004 (2004)
23. Andrejko, A., Barla, M., Bielikova, M.: Ontology-based user modeling for web-based information systems. Advances in Information Systems Development, pp. 457-468. Springer (2007)
24. Gouardères, G., Conté, E., Mansour, S., Razmerita, L.: Ontology based user modeling for personalization of grid learning services. 1st International ELeGI Conference on Advanced Technology for Enhanced Learning 8 (2005)
25. Jiang, X., Tan, A.-H.: Learning and inferencing in user ontology for personalized Semantic Web search. Information sciences 179, 2794-2808 (2009)
26. Gobbi, C., Hsuan, J.: Collaborative purchasing of complex technologies in healthcare: Implications for alignment strategies. International Journal of Operations & Production Management 35, 430-455 (2015)
27. Poon, C., Hung, K., Eren, H., Webster, J.: mHealth: Intelligent closed-loop solutions for personalized healthcare. Telehealth and mobile health, pp. 145-160. CRC Press (2015)
28. Spruit, M.A., Singh, S.J., Garvey, C., ZuWallack, R., Nici, L., Rochester, C., Hill, K., Holland, A.E., Lareau, S.C., Man, W.D.-C.: An official American Thoracic Society/European Respiratory Society statement: key concepts and advances in pulmonary

rehabilitation. American journal of respiratory and critical care medicine 188, e13-e64 (2013)
29. Benzo, R.P., Abascal-Bolado, B., Dulohery, M.M.: Self-management and quality of life in chronic obstructive pulmonary disease (COPD): The mediating effects of positive affect. Patient education and counseling 99, 617-623 (2016)
30. Warwick, M., Gallagher, R., Chenoweth, L., Stein-Parbury, J.: Self-management and symptom monitoring among older adults with chronic obstructive pulmonary disease. Journal of advanced nursing 66, 784-793 (2010)
31. Banos, O., Nugent, C.: E-coaching for health. Computer 51, 12-15 (2018)
32. Labaki, W.W., Han, M.K.: Improving detection of early chronic obstructive pulmonary disease. Annals of the American Thoracic Society 15, S243-S248 (2018)
33. Spathis, D., Vlamos, P.: Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. Health informatics journal 25, 811-827 (2019)
34. Agusti, A., Faner, R.: When Harry Met Sally, or When Machine Learning Met Chronic Obstructive Pulmonary Disease. American Thoracic Society (2020)
35. Gurbeta, L., Badnjevic, A., Maksimovic, M., Omanovic-Miklicanin, E., Sejdic, E.: A telehealth system for automated diagnosis of asthma and chronical obstructive pulmonary disease. Journal of the American Medical Informatics Association 25, 1213-1217 (2018)
36. Badnjevic, A., Gurbeta, L., Custovic, E.: An expert diagnostic system to automatically identify asthma and chronic obstructive pulmonary disease in clinical settings. Scientific reports 8, 1-9 (2018)
37. Badnjevic, A., Cifrek, M., Koruga, D., Osmankovic, D.: Neuro-fuzzy classification of asthma and chronic obstructive pulmonary disease. BMC medical informatics and decision making 15, S1 (2015)
38. Zarrin, P.S., Roeckendorf, N., Wenger, C.: In-vitro classification of saliva samples of COPD patients and healthy controls using machine learning tools. IEEE Access 8, 168053-168060 (2020)
39. Grua, E.M., Hoogendoorn, M.: Exploring clustering techniques for effective reinforcement learning based personalization for health and wellbeing. In: 2018 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 813-820. IEEE, (2018)
40. el Hassouni, A., Hoogendoorn, M., van Otterlo, M., Barbaro, E.: Personalization of health interventions using cluster-based reinforcement learning. In: International Conference on Principles and Practice of Multi-Agent Systems, pp. 467-475. Springer, (2018)
41. Sá-Sousa, A., Pereira, A.M., Almeida, R., Araújo, L., Couto, M., Jacinto, T., Freitas, A., Bousquet, J., Fonseca, J.A.: Adult Asthma Scores—Development and Validation of Multivariable Scores to Identify Asthma in Surveys. The Journal of Allergy and Clinical Immunology: In Practice 7, 183-190. e186 (2019)
42. Gower, J.C.: A general coefficient of similarity and some of its properties. Biometrics 857-871 (1971)
43. Gagolewski, M., Bartoszuk, M., Cena, A.: Genie: A new, fast, and outlier-resistant hierarchical clustering algorithm. Information Sciences 363, 8-23 (2016)
44. Durrani, Q.S.: Cognitive modeling: a domain independent user modeling. Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on 1, 217-220 (1997)