



Diagnosing asthma and chronic obstructive pulmonary disease with machine learning

Dimitris Spathis

Ionian University, Greece

Aristotle University of Thessaloniki, Greece

Panayiotis Vlamos

Ionian University, Greece

Abstract

This study examines the clinical decision support systems in healthcare, in particular about the prevention, diagnosis and treatment of respiratory diseases, such as Asthma and chronic obstructive pulmonary disease. The empirical pulmonology study of a representative sample ($n=132$) attempts to identify the major factors that contribute to the diagnosis of these diseases. Machine learning results show that in chronic obstructive pulmonary disease's case, Random Forest classifier outperforms other techniques with 97.7 per cent precision, while the most prominent attributes for diagnosis are smoking, forced expiratory volume 1, age and forced vital capacity. In asthma's case, the best precision, 80.3 per cent, is achieved again with the Random Forest classifier, while the most prominent attribute is MEF2575.

Keywords

asthma, clinical decision making, chronic obstructive pulmonary disease, machine learning, respiratory diseases

Introduction

With the increasing accumulation of big data, there are new opportunities and challenges related to the understanding, analysis and knowledge extraction. Simple statistical analysis of past decades is probably not enough in the era that everything is measurable. In this environment, the healthcare sector can benefit from identifying common patterns in different patients. But why did we choose to deal with asthma and chronic obstructive pulmonary disease (COPD)?

Corresponding author:

Dimitris Spathis, Department of Informatics, Ionian University, 7 Tsirigoti Square, Corfu 491 00, Greece.

Email: sdimitris@csd.auth.gr

Asthma is a chronic disease of the airways. The airways – or else *bronchi* – are tubes which carry air to the lungs. When an asthmatic inhales a stimulus from the environment, bronchospasm and restriction of air through them are induced. Patients with asthma have a hereditary predisposition, but symptoms appear after exposure to stimuli such as viral infections, and allergens (dust, pollen, animal dander, etc.). COPD refers to diseases associated with chronic bronchitis and emphysema, which have in common the respiratory airways obstruction, leading to daily dyspnoea. The main factor of the progressive airway obstruction is smoking. Every year, about 300 million patients are diagnosed with asthma, which causes a total of 250,000 deaths.¹ COPD occurs in 330 million patients worldwide, causing about 3 million deaths.² Consequently, the need to develop tools for early prediction and diagnosis of respiratory diseases arises.

This study examines the factors that characterise the diagnosis of asthma and COPD, with machine learning techniques. Machine learning is an artificial intelligence technique which allows us to learn from available data and make predictions for unseen conditions. The measurement data employed in the training of the algorithms were collected from chest-doctor visits. The long-term aim of this work is to be embedded in clinical decision support systems (CDSSs) for real-time management of asthma and COPD patients. Our work is in accord with published research in this field. In particular, Prosperi et al.³ compared linear and non-linear models in predicting asthma and eczema. Our methodology follows the same route, substituting eczema with COPD. They suggest that ‘more usefully-complex modelling is the key to a better understanding of disease mechanisms and personalised healthcare’. This is an interesting comment because we should test our systems not only in terms of performance but also in terms of complexity and interpretability. Our experiments take the above comment into consideration, testing non-linear and linear algorithms, providing interpretability with feature ranking. Also, we find many overlaps between the significant factors of diagnosing asthma and COPD as well, suggesting that we may model them together.

The remainder of this article is organised as follows. Section ‘Related work’ discusses the previous literature related to the problem of extracting the most important features from respiratory diseases, along with our research questions. The methodology followed is presented in section ‘Methodology’ dealing with data sources and tools used. Section ‘Results’ describes the statistical and machine learning results, along with a study of variable significance. Section ‘Discussion and conclusion’ provides some final conclusions, limitations and directions for future work.

Related work

The literature of this study is multifaceted, with applications in the fields of decision making, healthcare, and artificial intelligence. We chose to present some illustrative papers that resemble our methodology or present findings that helped us make some choices on our methodology. All papers are related to the leverage of computational methods regarding the diagnosis of respiratory diseases. We indicate the different variables that are used compared to this study. The thematic areas on which the literature review is divided are three: *CDSSs assessment*, *machine learning* and *clinical trials*. Clearly, there are many overlaps between the categories, but this separation is made for a better presentation of the work.

CDSSs assessment

A CDSS is any computer program designed to help healthcare professionals to make clinical decisions.⁴ Below, we present some relevant CDSSs related to the diagnosis of asthma and COPD.

An expert system based on rules was proposed by Khalid et al.⁵ and was created to cope with the uncertainty of suspected asthma. Their sample was 50 patients and the rules used were 96. Five

of their variables are used in this work too. The purpose of the study was to evaluate a CDSS against the judgment of the physician. The findings show that the area under the curve (AUC) for the system is at 0.9, while the doctor's is at 0.8. AUC area is a statistical measure that equals to a classifier's probability to classify a random positive value higher than a corresponding negative.

Children hospital visits constituted the dataset that was examined in Hoeksema et al.,⁶ who based their system on these data. The rules came from the national US standards. The sample comprises 1199 visits. The statistical methods chosen were the paired t-test and chi-square test. The different variables than ours were the 'reduction of the activities' of patients, hospitalisation due to asthma and the side effects of such medication. Research using the system showed that doctors agreed with the system 3/4 for mild cases, 1/3 for bold actions and 1/3 for those requiring action steps. The system responded with relative accuracy for assessment but not for therapy suggestion.

Another system based on national guidelines was that of Bi and Abraham,⁷ which was web based. The researchers, in this case, concluded that these systems, although they respond relatively well, have not been adopted widely because the implementation of directives and rules requires *continuous changes*.

Moving on to a meta-analysis, Matui et al.⁸ researched the adoption of systems in the last 20 years. After analysing 5787 investigations, they examined eight different experiments. The backbone of their findings was that the systems were inefficient, with low usage. When used, their advice was not followed. But when used properly, the systems improved prescribing rates and long-term clinical outcomes.

The difference between system adoption in urban and suburban areas was the subject of Bell et al.,⁹ in which they reviewed the adoption of a system in 12 medical centres for 1 year. The sample was 19,450 children between 0 and 18 years. The survey involved socioeconomic features including items such as origin (race) and type of insurance. Upon the system adoption, the findings showed an increase in the number of prescriptions for mild cases (6%) and spirometry rates (3%) in urban areas. At suburban areas, 14 and 6 per cent, respectively.

A similar meta-analysis of Fathima et al.¹⁰ involved experiments published in the last decade about CDSS assessment. The researchers concluded that the use of the systems improved the treatment of asthma and COPD in 3/4 of the cases, whereas almost half had a statistically significant improvement ($p < 0.05$) of the clinical results.

In the early years of the industry, Kuilboer et al.¹¹ created the 'Asthmacritic', a system that does not aim in deciding whether someone has the disease, but the next step, that of medicine and special treatment suggestion. After training, the system identified only 8.5 per cent of patients. During the research, the system made 255,664 recommendations to patients (approximately 3.4 per patient visit).

Machine learning

Many studies in literature, along with ours, focus on training systems with respiratory data in order to extract the most prominent features for diagnosis and prediction.

Machine learning algorithms for the diagnosis of asthma in expert systems were investigated by Prasad et al.¹² The algorithms compared were as follows: auto-associative memory neural networks (AMNN), Bayesian networks (BN), ID3 and C4.5, and the sample included 100 patients. The different variables than ours were the control before and after exercise and swelling in the legs. Findings show that Neural Networks outperform the other methods.

Asthma prediction in later life was the subject of the study by Pescatore et al.¹³ They examined children who developed asthma 5 years after the visit to the doctor for wheezing or coughing, compared to the automatic prediction. The different variables from ours were the presence of eczema

and the family history of asthma or bronchitis. The prediction method was the Logistic Regression on a sample of 1226 children. The 345 (28%) of them developed asthma 5 years after. The AUC score of the tool was 0.74.

A comparison between classifiers (artificial neural network (ANN), support vector machine (SVM), and Gaussian process) and expert-based BN (vs an automatically constructed one) was attempted by Dexheimer et al.¹⁴ Every classifier performed slightly worse than the expert-based model; BN achieved the highest accuracy.

Much alike our own methodology, Badnjevic et al.¹⁵ present a system for classification of asthma and COPD based on fuzzy rules and Neural Networks. The system was trained on 455 patients and achieved accuracy over 99 per cent in all cases. Their feature extraction was based on Global Initiative for Asthma (GINA) and Global Initiative for chronic Obstructive Lung Disease (GOLD) guidelines, which provided the fuzzy rules for the Neural Networks. This kind of pre-processing might have contributed to the almost perfect accuracy. In comparison to our study, they focused only on Neural Networks while we offer a thorough algorithmic machine learning comparison accompanied by feature significance ranking, without being based on data pre-processing. Also, while they analyse oscillometry data unlike us, our data contain demographic profile, medical and special lung measurements, habits and associated symptoms, attempting to look beyond only respiratory metrics. In comparison with oscillometry, standard spirometry continues to be the mainstay in the clinical assessment of lung function in children and adults.¹⁶ Oscillometry instruments are non-portable and more expensive compared to spirometers, only available in some major hospitals. Besides, our study focuses on variables that are easy to obtain towards the potential self-assessment of asthma and COPD; spirometers and oximeters are portable, cost below US\$600 and they are getting gradually integrated in wearables, smartphones and smart-watches. Mohktar et al.¹⁷ attempt to predict the risk of exacerbation in patients with COPD using home tele-health measurement data. The proposed algorithm, which uses a classification and regression tree (CART), has been validated with patients living at home. The algorithm can classify data into either a 'low-risk' or 'high-risk' category with 71.8 per cent accuracy and was able to detect a 'high-risk' condition 1 day prior to patients actually being observed as having a worsening in their COPD condition.

Similar to our setting, Himes¹⁸ identified clinical factors that modulate the risk of progression to COPD among asthma patients using data extracted from electronic medical records. A BN model composed of age, sex, race, smoking history, and eight co-morbidity variables is able to predict COPD in the independent set of patients with an accuracy of 83.3 per cent.

An unsupervised approach on this field was attempted by Fitzpatrick et al.¹⁹ Their experiment included the grouping (clustering) of 12 variables from 161 children in 5 hospitals for the Severe Asthma Research Program (SARP). The clustering method was the agglomerative hierarchical clustering, a bottom-up approach where each observation starts in its own cluster and gradually pairs of clusters are merged as we climb the hierarchy. The results of clustering showed four different clusters: (1) relatively good lung function and low atopy; (2) relatively low lung function, more atopy, increased symptoms and medication use; (3) greater co-incidence of diseases, increased bronchial reaction and lower lung function and (4) lower lung function, more symptoms and medication use. Significant clustering variables were the duration of asthma, the number of drugs and the threshold of pulmonary function. Prosperi et al.³ compared linear and non-linear machine learning methods such as Logistic Regression and Random Forests for predicting asthma and eczema in 554 adults. The findings show that the non-linear models show better sensitivity and specificity than other methods, more for asthma and wheezing and less for eczema. It is confirmed that the sensitivity to allergens and lung function characterises asthma better combined than separately. The Random Forest algorithm outperformed the others, verifying our own results as well in both cases.

Clinical trials

This section contains studies about experiments on respiratory diseases that did not employ a CDSS or machine learning, but most of the times researched more on the diseases per se with statistical inference.

Another study that deals with later life prediction of asthma is by Toelle et al.,²⁰ in which they tried to identify the factors that determine asthma in the later adult life of children. Out of 718 children aged 8–10 years, 565 (80%) were examined 15–17 years later. The statistical method used was the multivariate likelihood ratios (LRs). This measure involves the concepts of sensitivity and specificity. They recorded as separate variables the family history in respiratory disease from the mother and the father. Children's features that predict asthma symptoms in adulthood are low spirometry, hypersensitivity of the airways, recent whistle and the female sex (perhaps due to sampling). Children developing the above-mentioned five characteristics have an overall likelihood ratio of 36.9 for adult asthma existence.

At a little different viewpoint regarding the relation between the two under-review diseases, Soriano et al.²¹ studied the co-morbidity of asthma and COPD with other diseases, comparing the risk in the general population. The data were obtained from the UK General Practice Research Database. The sample was divided in patients with COPD ($n=2.699$) and asthma ($n=7.931$), which were diagnosed in 1998. The study showed that COPD co-occurs through the patients' lifetime along with angina, cataracts, bone fracture, osteoporosis, pneumonia and respiratory infections. However, asthma showed lower co-occurrence of diseases perhaps because of the younger age distribution sample, apart from respiratory infections and fractures.

Research questions

By studying the literature in these fields, we observe that the researchers have not reached a consensus for a generalised model of diagnosis and prediction of respiratory diseases. Researchers in the field of respiratory disease diagnosis have not decided on some common benchmarks in order to measure our techniques against them. From computational side, we cannot decide on the best algorithm for diagnosis because the available variables vary depending on the problem. Upon the above observations, this study attempts to answer the following research questions, by providing a comparison of some machine learning algorithms. The research questions we are attempting to answer are the following:

RQ1: What are the most important factors to diagnose asthma with machine learning methods, based on our sample?

RQ2: What are the most important factors to diagnose COPD with machine learning methods, based on our sample?

RQ3: Which machine learning algorithm achieves the highest precision in order to diagnose asthma and COPD?

Methodology

We are now able to structure the methodology of this empirical study. In order to examine the most dominant features of diagnosis and prediction of the two respiratory diseases, we analysed a dataset provided by a pulmonologist doctor. The sample contains 132 entries corresponding to unique patients who visited the clinic in a suburb of Thessaloniki, Greece, during the biennium 2014–2015. Each patient record describes 22 different values about the demographic profile, medical and

special lung measurements, habits and associated symptoms and, finally, the dependent variables: if one suffers from asthma or COPD.

The pulmonary measurements come from the spirometer, a device which estimates the volume of the air inhaled and exhaled by time. General medical measurements, such as oximetry and pulse, come from the oximeter. The remaining symptoms are typical questions asked by the physician to the patient. Also, we removed the variables Patient Name and Inhaler from our further analysis as being non-informative. Inhaler was used in Badnjevic et al.¹⁵ as a bronchodilatation medication to measure the lung reaction before and after the examination. In our dataset, it just indicates that the patient has already taken that medication, so it has no predictive value of diagnosis. We should note that we perform two different experiments for each disease (asthma, COPD). Table 1 provides a detailed view of each variable.

After the necessary cleaning and anonymisation of data, the sample was tested with the SPSS software. We present the descriptive statistics of each variable, correlations between the dependent and independent variables and a first prediction approach using Logistic Regression.

Besides the statistical analysis, the Python library ‘*Sklearn*’²² was used in order to apply machine learning algorithms for classification. Under the methodology of cross validation and hyperparameter optimisation, we tested multiple algorithms from different technical fields. We decided to choose a representative portion thereof used in the literature, attempting to cover a wide spectrum of classifiers’ categories such as linear, non-linear, kernel-based, trees and probabilistic ones. Decision Trees were used by Mohktar et al.,¹⁷ Prasad et al.¹² and Prosperi et al.³ Neural Networks were employed by Prasad et al.,¹² Dexheimer et al.¹⁴ and Badnjevic et al.¹⁵ Logistic Regression was compared by Prosperi et al.³ and Pescatore et al.¹³ Kernel-based methods such as SVMs were compared by Dexheimer et al.¹⁴ Also, Random Forest was examined by Prosperi et al.³ Finally, the final list of algorithms with which we deal is the following: Naive Bayes, Logistic Regression, Neural Networks, SVMs, K-Nearest Neighbour, Decision Trees and Random Forest.

Results

Following the above methodology, this section analyses the results of the dataset run through statistical and algorithmic models. The sample contains 132 entries corresponding to unique patients and 22 variables.

Statistical analysis

Tables 2 and 3 present the statistics of quantitative and qualitative variables. Along with the assistance of the doctor, we observe that the quantitative statistics seem normal; the spectrum of spirometry forced vital capacity (FVC) values ranges from 32 per cent, which indicates very poor airway performance, to 117 per cent, a totally healthy performance. The rest of the values follow the same pattern. At qualitative statistics, we observe that our sample contains more women than men (65% women) and the majority of patients are healthy; 45 per cent of them have asthma, while just 27 per cent have COPD.

The crosstab visualisation gives us a clear outline of the distribution of the dependent variable upon each independent variable of the dataset. In Figure 1, we observe that patients who have asthma display more frequently symptoms of chest pain, wheeze, use inhalers, perform lower at spirometry, suffer from breath shortness and cough, are females and belong to young age groups.

However, in Figure 2, patients who suffer from COPD are more likely to smoke, perform badly at spirometry, suffer more frequently from a cough and belong to older age groups. The observations of Figures 1 and 2 are considered normal by our doctor, which qualitatively validates our dataset.

Table 1. Variables used in our study.

No.	Attributes	Value	Comments
1	Age group	Numerical	Age groups are: 0–10 years old: 1 11–20 years old: 2 21–30 years old: 3 31–40 years old: 4 41–50 years old: 5 51–60 years old: 6 61–70 years old: 7 71–80 years old: 8 81–90 years old: 9
2	Sex	M/F	
3	Oxygen saturation	Numerical	The fraction of oxygen-saturated haemoglobin relative to total haemoglobin in the blood. Normal blood oxygen levels in humans are considered 95%–100%
4	Pulse	Numerical	
5	Cough	Yes, no	
6	Breath shortness	Yes, no	
7	FEV1	Numerical	Forced expiratory volume. FEV1 is the volume of air that can forcibly be blown out in 1 second, after full inspiration
8	FVC	Numerical	Forced vital capacity (FVC) is the volume of air that can forcibly be blown out after full inspiration
9	TIFF	Numerical	Tiffeneau–Pinelli index = FEV1/FVC
10	PEF	Numerical	Peak expiratory flow (PEF) is the maximal flow (or speed) achieved during the maximally forced expiration initiated at full inspiration
11	MEF25	Numerical	Maximal expiratory flow at 25% of the FVC curve
12	MEF50	Numerical	
13	MEF75	Numerical	
14	MEF2575	Numerical	
15	Inhaler	Yes, no	
16	Smoker	Yes, no	
17	Wheeze	Yes, no	
18	Spit	Yes, no	
19	Chest pain	Yes, no	
20	Patient name	Char	
21	Asthma	Yes, no	Dependent variable
22	COPD	Yes, no	Dependent variable

Patient names' initials only, due to privacy reasons (N = 132).

FEV1: forced expiratory volume; FVC: forced vital capacity; TIFF (Tiffeneau–Pinelli index) = FEV1/FVC; PEF: peak expiratory flow; MEF: maximal expiratory flow at 25/50/75 per cent of the FVC curve; COPD: chronic obstructive pulmonary disease.

Machine learning analysis

After statistical analysis, the sample is examined with machine learning algorithms. We selected the 20 independent variables and trained them against the two dependent variables (asthma and COPD) in two different experiments.

Table 2. Descriptive statistics of the quantitative variables.

	Mean	Standard error	Min	Max	Range	Standard deviation	Variance
Oxygen saturation	96.93	0.122	93	99	6	1399	1957
Pulse	82.96	1.109	50	125	75	12,740	162,296
FEV1	84.51	1.561	32	123	91	17,934	321,611
FVC	77.68	1.507	32	117	85	17,317	299,867
TIFF	113.62	0.925	70	140	70	10,628	112,955
PEF	80.26	1.711	28	128	100	19,661	386,559
MEF25	80.55	1.979	30	139	109	22,736	516,906
MEF50	84.79	2.607	23	168	145	29,949	896,947
MEF75	104.24	3.425	25	256	231	39,352	1548,597
MEF2575	92.80	2.673	27	188	161	30,710	943,106

FEV1: forced expiratory volume; FVC: forced vital capacity; TIFF (Tiffeneau–Pinelli index) = FEV1/FVC; PEF: peak expiratory flow; MEF: maximal expiratory flow at 25/50/75 per cent of the FVC curve.

Table 3. Descriptive statistics of the qualitative variables.

	Value	Frequency	Percentage
Age group (years)	1	5	3.8
	2	20	15.2
	3	20	15.2
	4	11	8.3
	5	26	19.7
	6	29	22.0
	7	13	9.8
	8	4	3.0
	9	4	3.0
Sex	F	86	65.2
	M	46	34.8
Cough	0	36	27.3
	1	96	72.7
Breath shortness	0	68	51.5
	1	64	48.5
Inhaler	0	105	79.5
	1	27	20.5
Smoker	0	91	68.9
	1	41	31.1
Wheeze	0	103	78
	1	29	22
Spit	0	75	55.3
	1	59	44.7
Chest pain	0	109	82.6
	1	23	17.4
Asthma	0	72	54.5
	1	60	45.5
COPD	0	96	72.7
	1	36	27.3

COPD: chronic obstructive pulmonary disease.

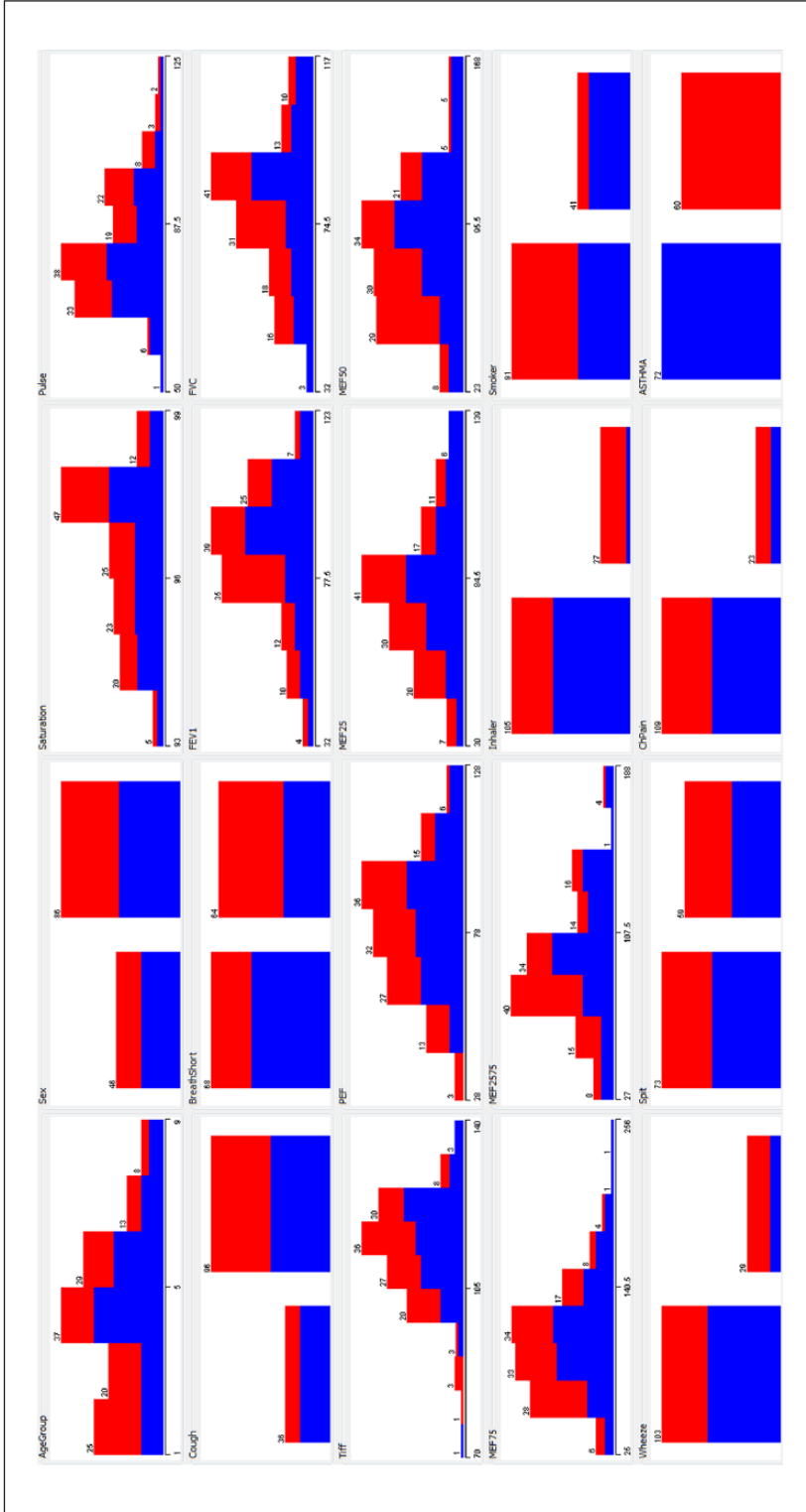


Figure 1. Asthma Crosstab visualisation. The crosstab visualisation gives us a clear outline of the distribution of the dependent variable upon each independent variable of the dataset. Light colour (red) signifies the patients with asthma, while dark (blue) the healthy ones.

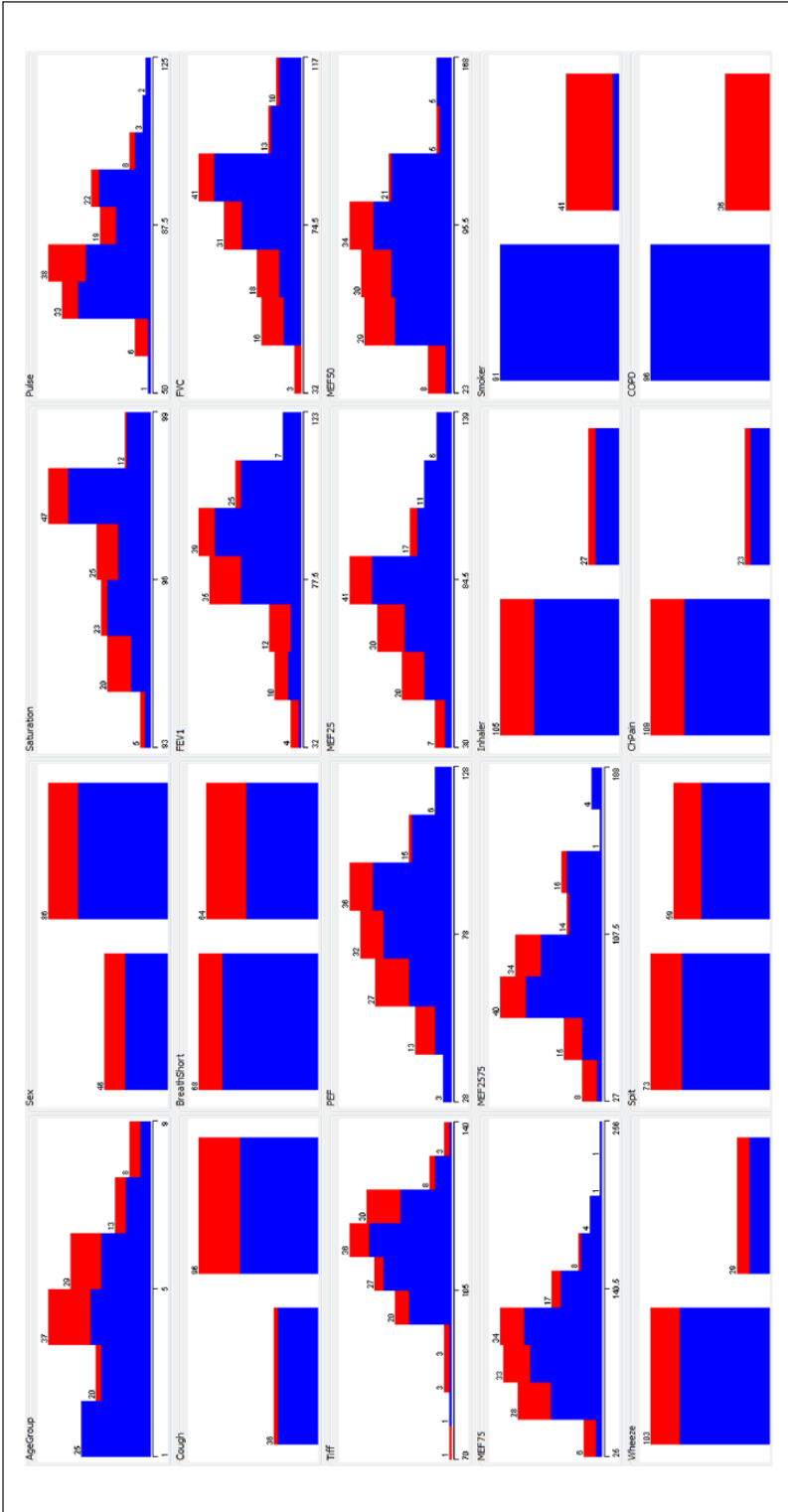


Figure 2. COPD crosstab visualisation. The crosstab visualisation gives us a clear outline of the distribution of the dependent variable upon each independent variable of the dataset. Light colour (red) signifies the patients with COPD, while dark (blue) the healthy ones.

Table 4. Comparison of machine learning algorithms for asthma diagnosis.

Classifiers	Training		20% test validation		Training		20% test validation	
	10-fold CV precision		Precision	F1	5-fold CV precision		Precision	F1
	Naive Bayes	73.4	82	81	70.4	82	81	
Logistic regression	71.2	67	67	72.7	67	67		
Neural Network	66.6	37	36	66.6	72	64		
SVM	71.2	70	70	70.4	70	70		
K-Nearest Neighbour	68.9	64	62	67.4	60	57		
Decision Tree	71.2	67	66	68.9	63	63		
Random Forest	80.3	83	81	79.5	83	81		

SVM: support vector machine; CV: cross validation.

This table presents the precision of the cross-validated hyper-parameter optimisation, performed with 5 and 10 folds. After the grid search, we obtained the optimal hyper-parameters, split the data into 20 per cent test and 80 per cent train and predicted on the test. We report precision and F1 score.

In order for each algorithm to be able to process the data, we pre-processed the categorical variables of sex (male–female), transforming them to numerical with a technique called One Hot Encoding.

We evaluate our algorithms with precision and f-measure scores. Precision (P) is the ratio of True Positives (TP) to True Positives (TP)+False Positives (FP). F-measure (or F1 score) uses Precision (P) and Recall (R), with recall being the ratio of True Positives (TP) to True Positives (TP)+False Negatives (FN), so that F-measure is the harmonic mean thereof. Precision is one of the most common metrics in machine learning literature and is preferred to accuracy because when there is a large class imbalance, it may be desirable to select a model with a lower accuracy due to its greater predictive power on the problem

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{F-Measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Machine learning algorithms contain some hyper-parameters that should be fine-tuned. We performed a methodology called Grid Search, searching in the space of all possible parameters of the supported algorithms. The only algorithm that does not need this methodology is Naive Bayes,²³ being non-parametric. During Grid Search, we evaluate each model with the reported precision on 5- and 10-fold cross validation. These values are reported on the column Training in Tables 4 and 5.

After training, we keep the models with the highest precision as well as their hyper-parameter values. In order to test them in an unexpected real-world scenario, we split the dataset into 20–80 per cent, with 20 per cent being the test set and 80 per cent the train set. Please note that we do

Table 5. Comparison of machine learning algorithms for COPD diagnosis.

COPD						
Classifiers	Training	20% test validation		Training	20% test validation	
	10-fold CV precision	Precision	F1	5-fold CV precision	Precision	F1
Naive Bayes	96.9	97	96	96.9	97	96
Logistic regression	96.2	97	96	95.4	97	96
Neural Network	81.8	88	83	82.5	89	89
SVM	93.9	97	96	92.4	97	96
K-Nearest Neighbour	76.5	50	58	75.7	63	64
Decision Tree	96.2	97	96	94.6	100	100
Random Forest	97.7	100	100	96.9	100	100

COPD: chronic obstructive pulmonary disease; SVM: support vector machine; CV: cross validation.

This table presents the precision of the cross-validated hyper-parameter optimisation, performed with 5 and 10 folds. After the grid search, we obtained the optimal hyper-parameters, split the data into 20 per cent test and 80 per cent train and predicted on the test. We report precision and F1 score.

not use the already trained models for prediction; we only take their optimal parameters. We train the algorithms passing them the best hyper parameters and test them with the 20 per cent unseen data. These values (precision and F1 score) are reported on the column ‘Test Validation’ in Tables 4 and 5.

The obvious point drawn by Tables 4 and 5 is that the predictive power of COPD is stronger than in asthma, with precision difference around 15 per cent between the two. Nevertheless, the results are quite high and allow us to use the models in forecasting.

In asthma’s case, we observe the highest precision (80.3%) by the Random Forest classifier and 10-fold cross validation. The validation test-set achieves 83 per cent precision. Random Forests,²⁴ as a combined solution (ensemble), solve the problem of over-fitting in conventional Decision Trees and seem to be the perfect solution for interpretability and predictive power. Their models are easily interpreted even by non-experts and their results are very accurate. Our optimal hyper-parameters after the Grid Search were the following: minimum number of samples required to be at a leaf node: 1; quality of a split criterion: entropy; number of trees in the forest: 100; features to consider when looking for the best split: $\sqrt{\text{features}}$; maximum depth of the tree: 100; and minimum number of samples required to split an internal node: 2.

Apart from Random Forest, Naive Bayes comes second with 73.4 per cent in 10-fold cross validation, while Logistic Regression is third with 72.7 per cent in 5-fold cross validation.

In COPD’s case, the highest precision is achieved again by Random Forests (97.7%) and 10-fold cross validation, with the optimal hyper-parameters being minimum number of samples required to be at a leaf node: 1, quality of a split criterion: entropy, number of trees in the forest: 50, features to consider when looking for the best split: $\log 2$, maximum depth of the tree: 5 and minimum number of samples required to split an internal node: 2. Apart from Random Forests, again Naive Bayes (96.9%) and Logistic Regression (96.2%) produce high precisions, both with 10-fold cross validation.

Comparing the two cross-validation techniques (with 5 and 10 folds), in both cases, 10 fold performs better with the exception of Logistic for asthma and Neural Network for COPD, when 5 fold performed better.

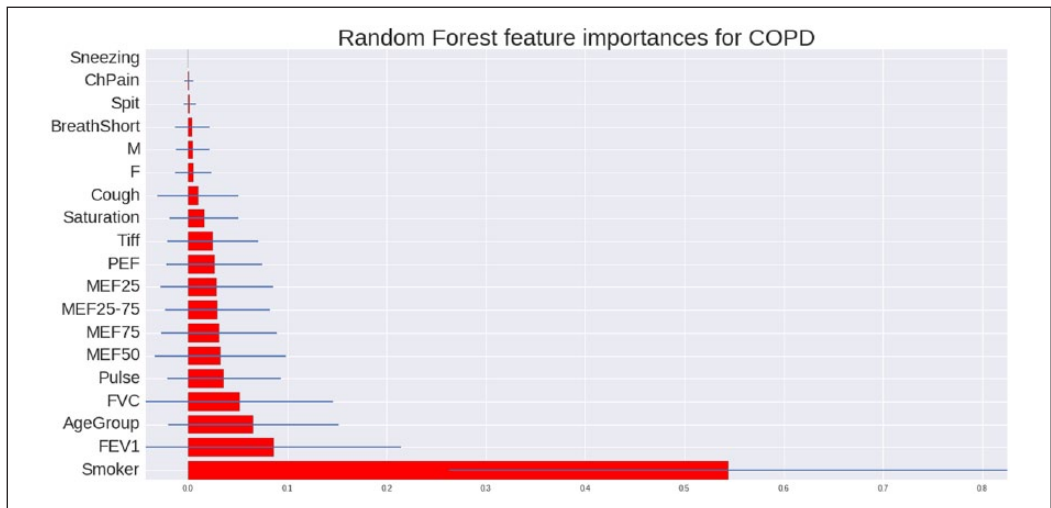


Figure 3. Feature importance for COPD, extracted from the Random Forest with the best accuracy. The red bars are the feature importance of the forest, along with their inter-trees variability.

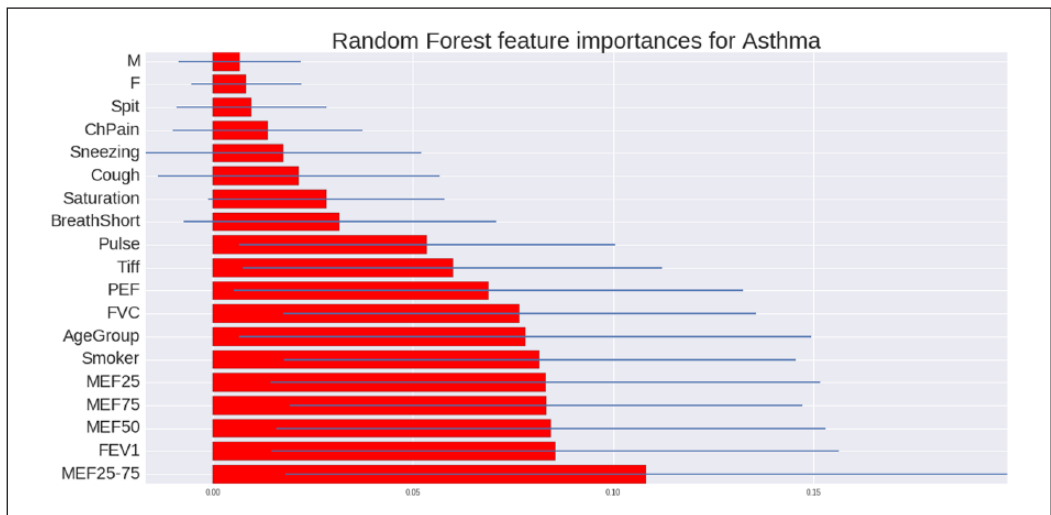


Figure 4. Feature importance for Asthma, extracted from the Random Forest with the best accuracy. The red bars are the feature importance of the forest, along with their inter-trees variability.

Since Random Forests outperform other machine learning techniques in both cases, we obtain their feature significances in order to be able to understand which features are important to them. In Figures 3 and 4, we graph each feature importance (red bar) along with its *inter-tree variability* (blue deviation line). We see that for both asthma and COPD, the feature ranking is similar with the results we obtain in Feature Ranking, which we discuss in the next section.

Table 6. Feature ranking for COPD and asthma.

Rank	COPD		Asthma	
	InfoGain	Correlation	InfoGain	Correlation
1	Smoker	Smoker***	MEF2575	Smoker*
2	FEV1	FEV1*	Age	MEF2575*
3	Age	FVC*	Smoker	MEF50*
4	FVC	MEF50*	Breath shortness	Age*
5	Cough	Age*	Wheeze	Wheeze*
6	Breath shortness	MEF25*	Cough	MEF25*
7	MEF50	MEF2575*	Chest pain	Breath shortness
8	Wheeze	PEF*	Sex	TIFF
9	MEF25	MEF75*	Spit	MEF75
10	Sex	Cough*	MEF50	PEF

FEV1: forced expiratory volume; FVC: forced vital capacity; TIFF (Tiffeneau–Pinelli index) = FEV1/FVC; PEF: peak expiratory flow; MEF: maximal expiratory flow at 25/50/75 per cent of the FVC curve.

Ranking of variables with information gain and Pearson correlation.

Correlation: (*) > 0.25, (**) > 0.50 and (***) > 0.75.

Feature ranking

The selection of attributes that determine the classification into one class or another is one of the most important decisions in a machine learning experiment. In our experiment, we are interested in identifying which feature influences the final classification more. Using two attribute selection methods, Correlations and Information Gain, we conclude in the ranking shown in Table 5. We restrict the ranking up to 10 features, in order to focus and identify the most important ones.

The correlation method (Pearson's) is among the most prevalent in the field of statistics on finding relationships between variables. InfoGain method calculates the entropy change from one state to another, receiving information according to the following formula

$$\text{InfoGain}(\text{Class, Attribute}) = H(\text{Class}) - H(\text{Class}|\text{Attribute})$$

In Decision Trees, comparing the entropy before and after the split, we obtain a measure of information gain or how much information we gained by doing the split using that particular feature.

Significant COPD variables. In Table 6, we observe that the two methods agree to the variables: *Smoker*, *FEV1*, *Age*, *FVC*, *Cough*, *MEF50* and *MEF25*. However, the position of their classification varies. We note that these seven variables encountered in both methods are related mostly to pulmonary spirometer measurements, demographics and symptoms. Of course, the most prevalent feature is smoking. We know from literature that COPD occurs in middle-aged and elders, especially in smokers. This is clearly reflected in the choice of variables. Older age and smoking history are confirmed by Himes et al.,¹⁸ with age being the best predictor if we do not know about the smoking habits. Also, Hamad et al.²⁵ found in their tele-health system that cough created significantly more alerts by those treated in the community, whereas a drop in oxygen saturation created significantly more alerts pre-hospitalisation.

Significant asthma variables. In asthma's case, the unanimity we observed in the case of COPD is absent. Asthma affects a larger population demographic, making it difficult to analyse and identify the most important factors. The common characteristics of the two compared methods are *Inhaler*, *MEF2575*, *Age*, *Smoker*, *Wheeze* and *Breath Shortness*. MEF2575 describes the maximal value of expiration rate in the middle of the FVC curve and the low metrics as predictors are confirmed by Fitzpatrick et al.¹⁹ and Prosperi et al.³ Additionally, this single pulmonary measurement occurs on both methods, unlike COPD. This, according to the physician, is interpreted as that the asthma is reversible during the life of the patient. There are periods of life that one can achieve 100 per cent spirometry measurements. Low values are mainly observed in periods of asthma attacks, which are relatively short. The symptoms, however, such as breath shortness and wheezing, appear regularly.

Discussion and conclusion

Our study examines the CDSSs in healthcare, in particular about the prevention, diagnosis and treatment of respiratory diseases, such as Asthma and COPD. The empirical study of a representative sample (*132 patients* and *22 attributes*) identified the major factors that contribute to the diagnosis of the above diseases. Our machine learning results show that in COPD's case, Random Forests outperform other techniques with 97.7 per cent precision, while the most prominent attributes for diagnosis are smoking, FEV1, age and FVC. In asthma's case, the best precision, 80.3 per cent, is achieved with the Random Forest classifier, while the most prominent attributes are MEF2575, smoking, age and wheeze. Regarding the machine learning methods, asthma is detected more accurately with Random Forest, a fact that is confirmed by Prosperi et al.³ too.

Other studies did not distinguish so much between different spirometer metrics or included one or two, while we include eight. We observe that despite being measured by the same device, there are big differences in their predictive power. For instance, FEV1, the volume of air that can forcibly be blown out in 1 s, is the most important spirometer factor for both cases, while Tiffeneau–Pinelli index (TIFF) significance – being the ratio of FEV1/FVC – is almost negligible. These observations would be valuable to pulmonologists.

The findings of our study, namely, the most important factors of asthma and COPD modelling for a prediction model, may assist doctors in their final assessment and monitoring of the patient. Apart from physicians, it should help IT and software engineers who develop expert decision support systems to evaluate different algorithms that provide the best classification of the above diseases.

As per the limitations of this study, we acknowledge that the size of the sample is not considered big for us to be able to generalise our predictions. However, the variety of the attributes/features allows us to describe each patient of our dataset thoroughly.

In fact, the novelty of our study lies on the leverage of many factors as attributes; from demographic data to spirometry rates, general medicine metrics and common symptoms. From our discussions with clinicians and doctors, we understand that the challenge on the adoption of expert systems on this field is a matter of user experience and standards. Especially on the latter, Bi and Abraham⁷ reached the same conclusion that continuous changes in national guidelines and technologies do not help the further integration. Another challenge to be researched is the implementation of predictive models that do not require *invasive* medicine or expensive devices; spirometry in our case. Would it be possible to predict respiratory diseases just from habits, demographic, socio-economic or air-pollution data? Mobile apps and quantified-self devices should surely contribute to that. That is the next question for big data analysis in order to democratise healthcare through technology.

Acknowledgements

The authors would like to thank editor Prof. Rob Procter and the anonymous reviewers for their constructive comments that helped to improve the manuscript. Also, the authors acknowledge the support of Antonia Ziaka, MD, and medical student Maria Spathi for collecting the data and interpreting the results. We would also like to thank Telefonica Research for providing computational resources during first author's fellowship.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Bateman ED, Hurd SS, Barnes PJ, et al. Global strategy for asthma management and prevention: GINA executive summary. *Eur Respir J* 2008; 31(1): 143–178.
2. Naghavi M, Wang H, Lozano R, et al. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 2015; 385(9963): 117–171.
3. Prosperi MC, Marinho S, Simpson A, et al. Predicting phenotypes of asthma and eczema with machine learning. *BMC Med Genom* 2014; 7(Suppl. 1): S7.
4. Musen MA, Middleton B and Greenes RA. Clinical decision-support systems. In: Shortliffe EH and Cimino JJ (eds) *Biomedical informatics*. London: Springer, 2014, pp. 643–674.
5. Khalid MS, Haque MA, Hossain MS, et al. A belief rule-based (BRB) decision support system for assessing clinical asthma suspicion. In: *Proceedings of Scandinavian conference on health informatics*, Grimstad, 21–22 August 2014, pp. 83–89. Linköping: Linköping University Electronic Press.
6. Hoeksema LJ, Bazzy-Asaad A, Lomotan EA, et al. Accuracy of a computerized clinical decision-support system for asthma assessment and management. *J Am Med Inform Assoc* 2011; 18(3): 243–250.
7. Bi J and Abraham A. An intelligent web-based decision support tool for enhancing asthma guideline adherence. In: *Proceedings of the 2nd ACM SIGHIT international health informatics symposium*, Miami, FL, 28–30 January 2012, pp. 51–60. New York: ACM.
8. Matui P, Wyatt JC, Pinnock H, et al. Computer decision support systems for asthma: a systematic review. *NPJ Prim Care Respir Med* 2014; 24: 14005.
9. Bell LM, Grundmeier R, Localio R, et al. Electronic health record-based decision support to improve asthma care: a cluster-randomized trial. *Pediatrics* 2010; 125(4): e770–e777.
10. Fathima M, Peiris D, Naik-Panvelkar P, et al. Effectiveness of computerized clinical decision support systems for asthma and chronic obstructive pulmonary disease in primary care: a systematic review. *BMC Pulm Med* 2014; 14(1): 189.
11. Kuilboer MM, Van Wijk MA, Mosseveld M, et al. Feasibility of AsthmaCritic, a decision-support system for asthma and COPD which generates patient-specific feedback on routinely recorded data in general practice. *Fam Pract* 2002; 19(5): 442–447.
12. Prasad BDCN, Prasad PK and Sagar Y. A comparative study of machine learning algorithms as expert systems in medical diagnosis (Asthma). In: Meghanathan N, Kaushik BK and Nagamalai D (eds) *Advances in computer science and information technology*. Berlin: Springer, 2011, pp. 570–576.
13. Pescatore AM, Dogaru CM, Duembgen L, et al. A simple asthma prediction tool for preschool children with wheeze or cough. *J Allergy Clin Immunol* 2014; 133(1): 111–118.
14. Dexheimer JW, Brown LE, Leegon J, et al. Comparing decision support methodologies for identifying asthma exacerbations. *Stud Health Technol Inform* 2007; 129(Pt 2): 880.

15. Badnjevic A, Cifrek M, Koruga D, et al. Neuro-fuzzy classification of asthma and chronic obstructive pulmonary disease. *BMC Med Inform Decis Mak* 2015; 15(Suppl. 3): S1.
16. Komarow HD, Myles IA, Uzzaman A, et al. Impulse oscillometry in the evaluation of diseases of the airways in children. *Ann Allergy Asthma Immunol* 2011; 106(3): 191–199.
17. Mohktar MS, Redmond SJ, Antoniadis NC, et al. Predicting the risk of exacerbation in patients with chronic obstructive pulmonary disease using home telehealth measurement data. *Artif Intell Med* 2015; 63(1): 51–59.
18. Himes BE, Dai Y, Kohane IS, et al. Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records. *J Am Med Inform Assoc* 2009; 16(3): 371–379.
19. Fitzpatrick AM, Teague WG, Meyers DA, et al. Heterogeneity of severe asthma in childhood: confirmation by cluster analysis of children in the National Institutes of Health/National Heart, Lung, and Blood Institute Severe Asthma Research Program. *J Allergy Clin Immunol* 2011; 127(2): 382–389.
20. Toelle BG, Xuan W, Peat JK, et al. Childhood factors that predict asthma in young adulthood. *Eur Respir J* 2004; 23(1): 66–70.
21. Soriano JB, Visick GT, Muellerova H, et al. Patterns of comorbidities in newly diagnosed COPD and asthma in primary care. *Chest* 2005; 128(4): 2099–2107.
22. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; 12: 2825–2830.
23. John GH and Langley P. Estimating continuous distributions in Bayesian classifiers. In: *Proceedings of the Eleventh conference on uncertainty in artificial intelligence*, Montreal, QC, Canada, 18–20 August 1995, pp. 338–345. Burlington, MA: Morgan Kaufmann Publishers Inc.
24. Breiman L. Random forests. *Mach Learn* 2001; 45(1): 5–32.
25. Hamad GA, Crooks M and Morice AH. The value of telehealth in the early detection of chronic obstructive pulmonary disease exacerbations: a prospective observational study. *Health Informatics J* 2015; 22(2): 406–413.