Rodríguez-Álvarez et al. J Fam Med Dis Prev 2016, 2:045

Volume 2 | Issue 4

ISSN: 2469-5793

Journal of
Family Medicine and Disease Prevention

Original Research Article: Open Access

Real-Data Comparison of Data Mining Methods in Early Detection of Chronic Obstructive Pulmonary Disease (COPD) in General Practice

Rodríguez-Álvarez Cristobalina[1]*, Rupérez Félix[2], González-Dávila Enrique[3], González-Martín Isidro[4], Castro Beatriz[4] and Arias Ángeles[1]

[1]Department of Preventive Medicine and Public Health, University of La Laguna, Canary Islands, Spain

[2]Department of Nursing, University of La Laguna, Spain

[3]Department of Mathematics, Statistics and Operations Research, University of La Laguna, Spain

[4]University Hospital of the Canary Islands, Spain

***Corresponding author:** *Ángeles Arias Rodríguez, Department of Preventive Medicine and Public Health, University of La Laguna, Canary Islands, Spain, Tel: +34-922-319369, E-mail: angarias@ull.es*

Summary

Background and aims: Health authorities have increased the attention given to chronic obstructive pulmonary disease (COPD) in recent years. Even so, under-diagnosis and late diagnosis rates remain high. The aim of this study was to determine which factors allow to discriminate between people with COPD and which do not, trying to provide a simple tool that can be used by primary health care personnel.

Methods: A cross-sectional epidemiological study was carried out on the island of Tenerife with a final sample of 402 individuals. Using five different methods of data mining, we assessed the ability of anthropometric variables and items of a questionnaire related to respiratory disease to discriminate between patients with and without COPD.

Results: Individually, dyspnoea when climbing one flight of steps was the item with a greater degree of discrimination. The five implemented methods show high reliability to detect patients with COPD. In particular, JRip decision rule has the best sensitivity (0.911) and the second best area under the receiver operating characteristic curve (0.932) of the methods evaluated. The number of variables included in this model was only five: dyspnoea when climbing one flight of steps, dyspnoea when walking, daily phlegm and cough for three months, and smoking intensity.

Conclusions: Introduced decision rules are a simple and easy tool to improve early detection of COPD in primary health care centres, enabling the detection of patients where a spirometry test should be performed. We propose a simple predictive model, with a number of easy to obtain items that have demonstrated capacity to identify patients with and without COPD. Performance characteristics suggest that our questionnaire could be very useful to enhance efficiency and diagnostic accuracy of current screening efforts using spirometry alone, so our model would be useful to improve the accuracy of early diagnosis of COPD in smokers with respiratory symptoms. Therefore, we consider that is a model that might be applied by the nurse specialist in primary healthcare, as a previous step of the spirometry intervention that confirms the diagnosis of the disease.

Keywords

Data mining, Decision rules, Predictive model, Chronic obstructive pulmonary disease, Primary health care

Introduction

Chronic obstructive pulmonary disease (COPD) is one of major cause of morbidity and mortality in developed countries [1]. In spite of the fact that this disease is narrowly tied to tobacco consumption and that the developed countries are adopted important campaigns for the prevention of it, the disease prevalence and mortality continue increasing worldwide. According to clinical forecasts, in 2020, this disease will be the fifth cause of disease and the third cause of mortality, worldwide. Despite of that reasons, this disease is receiving, in the last years, an increasing medical attention, nevertheless still it is relatively ignored by the population, by the public health and the governments [2].

COPD has a significate impact in the quality life of patient and in costs supported by health system. A severe form of COPD is the most common condition that requires hospitalization and substantially contributes to the economic related impact. This includes the excessive cost of all the medicines with medical prescription, attention in general medicine, emergency rooms and the episodes of hospitalization [3].

COPD is a complex, chronic and progressive disease characterized by the chronic inflammation and irreversible air flow obstruction, which involves structural changes in the lung. The principal symptoms are the difficulty in breathing, cough and expectoration. In the clinical presentation are different phenotypes, very heterogeneous, with prognostic and therapeutic clinical repercussions [4].

Though COPD is not a curable disease, to stop smoking is the most effective measure for prevention and to stop the progression. COPD's clinical diagnosis must think about every patient with a respiratory difficulty, chronic cough or high production of secretions and a history of exposition to risk factors of the diseases [5]. Different authors indicate as significant factors related to this condition: the masculine sex, age, consumption of tobacco (number of packages per year), cough, expectoration, difficulty in breathing and other respiratory symptoms [6-10].

ClinMed
International Library

Citation: Rodríguez-Álvarez C, Félix R, González-Dávila E, González-Martín I, Beatriz C, et al. (2016) Real-Data Comparison of Data Mining Methods in Early Detection of Chronic Obstructive Pulmonary Disease (COPD) in General Practice. J Fam Med Dis Prev 2:045

Received: September 28, 2016: **Accepted:** November 03, 2016: **Published:** November 07, 2016

Several studies indicate that to achieve a good control of COPD is essential to do a diagnosis in the first stages of the disease, as well as to adopt more appropriate preventive measures and to assure a systematic control and a good follow-up of the disease [2,11].

Therefore, the early detection and the diagnosis of the COOD play an important paper in the effective strategies of prevention. Nevertheless, there is not any general model that has been generalized in primary healthcare for this purpose. For that, it is important to provide an efficient and precise model to predict the population at risk of suffering COPD, to identify the not diagnosed people who require a diagnosed by spirometry.

Certain authors address the essential issue of how to implement a selection model to diagnose COPD in early stages [7,12]. Dirven, et al. [8,9] indicate that questionnaires could be conducted in primary care to detect respiratory health problems and, depending on their results, subsequently prescribe a spirometry leading to an accurate diagnosis. Clinicians and health service researchers are frequently interested in predicting patients' specific probabilities of adverse events (e.g. death, disease recurrence, post-operative complications and hospital readmission) [13]. Data mining has helped to predict under-diagnosed patients, as well as to identify and classify at-risk people in terms of health [14-16]. The aim of this study was to determine which factors allow to discriminate between people with COPD and which do not, trying to provide a simple tool that can be used by primary health care personnel. To answer this question, we have applied five different methods commonly used in data mining, two methods of decision trees, two of decision rules and one method of decision function. The results enable high efficiency prediction of COPD using a reduced number of factors which may be easily employed in the field of primary health care.

Methods

Design

To carry out the objectives of this work, a cross-sectional epidemiological study was conducted on the Island of Tenerife (Spain) during the period running from September 1, 2011 to December 31, 2012, in which individuals, smokers of both sexes, between 40 and 69 years of age were included.

We selected one third of the 37 health centers in Tenerife and stratify in the four geographical areas of health: Metropolitan, North, Southeast and Southwest, with proportional allocation to the centers of each, 16, 12, 4 y 5 respectively. Thus, a total of 12 centers were selected. The necessary permissions were obtained, as well as the collaboration of family doctors, nurses and staff from the centres. All centres had a spirometer model Datospir 120 (Sibel S.A.), which is the model available in primary care centres in Tenerife.

The total number of participants, 2,163 was determined using the total population between 40 and 69 years in Tenerife (Continuous Register 2010, 348,844 inhabitants), a limit proportion of COPD in smokers of 15%, a significance level of 5% and the accuracy in the estimating of 1.5%. These participants were randomly selected in the 12 centers included after a deal with proportional allocation based on the number of patients assigned to each center.

Inclusion criteria considered were: to be between 40 and 69 years of age, to have a positive history of smoking (current or ex-smoker), not having any previous test of respiratory diagnosis, and to be willing to collaborate and subsequently sign the informed consent. Thereafter, an appointment at the health centre was arranged with these people. The exclusion criteria were: to be out of range age, to be neither smoker nor previously smoker, patients with previous COPD diagnosis.

The appointments of the individuals were made by sending them a personalized letter containing a brief overview of the study and indicating they had been randomly selected. They were invited to participate in the study and informed that they would be telephoned in the coming days.

Affiliation data, age, sex and the values provided by the questionnaire of the European Coal and Steel Community (ECSC) on respiratory symptoms translated and validated in Spain [17] were collected from participants. This questionnaire includes different sections related to respiratory disease, such as the existence of cough and expectoration, dyspnoea, wheeze and chest oppression, among others. A smoking index in term of pack-years calculated with the information on the number of cigarettes smoked per day and the number of years the person has smoked was also collected.

All participants were previously instructed about the test to be performed. On each individual three spirometry were performed by a single skilled person, following the recommendations of the American Thoracic Society and always using the same type of spirometer already mentioned.

Lung function measurements included forced expiratory volume in 1 second (FEV1), forced vital capacity (FVC) and their ratio (FEV1/FVC). FEV1 and FVC were expressed in litres and as the percentage relative to the reference values for the Spanish population. According to the Spanish COPD guidelines and as proposed elsewhere for mass screening programs, we used pre-bronchodilator lung function to classify airflow limitation, defined by an FEV1/FVC ratio < 0.70 [18].

Data analysis and experimental configuration

All statistical analyses and prediction models were conducted in SPSS 21 for Windows (IBM SPSS Statistics, Chicago, IL, USA) and Weka 3.6.3. (Waikato Environment for Knowledge Analysis, GNU-GPL). No missing data occurred because the researcher was present at the taking of information. Different data mining methods were tested with the intention of obtaining a good model for predicting COPD. In particular, two methods of decision tree, J48 (version of C4.5 in Weka) and CART, two methods of decision rules, JRip and PART, and one method of decision function, the logistic regression (LR) [19] were used. To increase the predictive quality, we initially applied the wrapper-based approach in the variable selection process [20]. This allowed us to find a quasi-optimal set of variables associated with the data mining method which would then be applied. In all cases, a Genetic Search algorithm was used. Other search algorithms, as Best-First, Greedy Stepwise, Random or Exhaustive, were discarded for producing a very small set of variables or excessive computation time.

Area under the receiver operating characteristic curve (AUROC), sensitivity, 1- specificity (false positive rate), F-measure, and Cohen's Kappa are reported to assess the efficiency of selected models. These statistics are shown using both the total sample as training set, as well as after evaluating the model with 10-fold cross validation. Since in these studies the sample has a very high number of non-COPD compared with COPD and the validation test of disease is not too expensive, we used sensitivity as the primary criterion for the comparison of predictive power.

Results

Of 2.163 individuals previously selected, 18.7% (24.4% of men and 13.6% of women) met the inclusion criteria and went to their health centres on the day concerted (Figure 1). The sample was finally composed of 265 men and 147 women.

Of the 402 individuals analysed 45 (11.2%) were diagnosed as COPD. The percentages of non-COPD and COPD for each variable considered in this study are presented in table 1. Predictive power of individual characteristics for COPD, measured as AUROC and odds ratio, is also shown in that table. In particular, 14.6% of men had COPD compared with 5.3% of women (p = 0.005). The percentage of COPD increased as a function of age group, from 3.2% in those younger than 50 to 21.9% in those over 60 (p < 0.001). Similarly, the percentage of COPD increased the greater the smoking intensity (pack-years) from 1.1% in the group with < 15 pack-years to 19.7% in the group with more than 30 pack-years (p < 0.001). Of the items listed in the ECSC questionnaire, those relating to dyspnoea, cough, wheeze and phlegm stand out, among others. For example, 81.0%,

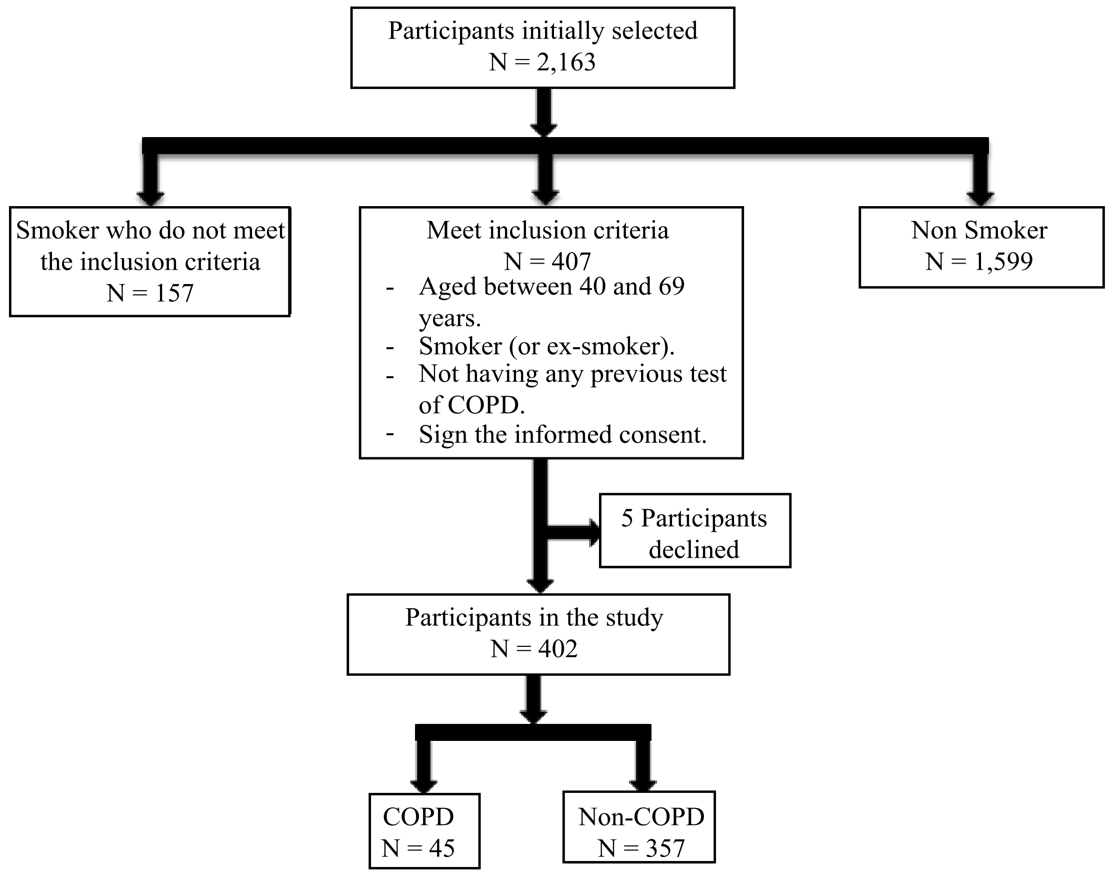
Participants initially selected
N = 2,163
Smoker who do not meet the inclusion criteria
N = 157
Meet inclusion criteria
N = 407
- Aged between 40 and 69 years.
- Smoker (or ex-smoker).
- Not having any previous test of COPD.
- Sign the informed consent.
Non Smoker
N = 1,599
5 Participants declined
Participants in the study
N = 402
COPD
N = 45
Non-COPD
N = 357

Figure 1: Flow diagram of participants.

68.9% and 87.5% of those with dyspnoea when walking (without other diseases), climbing one flight of steps and dressing or washing, respectively, had COPD, compared with 7.3%, 0.9% and 9.6% of those without such symptoms (all p < 0.001).

The best indicator of the risk of COPD is dyspnoea when climbing one flight of steps (OR = 249.05; AUROC = 0.94), followed by dyspnoea when walking on level ground, cough daily, cough in the morning and wheeze, all with an AUROC > 0.7.

Table 2 shows the resulting predictive models for COPD when combining the different characteristics observed in the five data mining methods used. As expected, all include dyspnoea when climbing one flight of steps. These models are obtained after the application of the wrapper-based approach in the variable selection process. The selected variables are listed in table 3. The number of variables included in the final models ranged from 3 in the J48 decision tree to 6 in the PART decision list and logistic regression, with 5 in the case of the CART decision tree and JRip rule.

As an example of using table 2, consider an individual 40 years of age who has dyspnoea when climbing one flight of steps and a family history of asthma. If we apply the CART decision tree method we must classify the individual as non-COPD since *"Dyspnoea when climbing one flight of steps = Yes", "Age group= < 50 years" and "Asthma family history = Yes"*. In particular, for the sample used, once applied the 10-fold cross validation, the 6 people who met these features were classified correctly, as shown in brackets (6/0) in that table. If we apply the J48 decision tree, we get that we must classify him as non-COPD since *"Dyspnoea when climbing one flight of steps = Yes" and "Age group= < 50 years"*. In this case, of the 21 people in the sample who met the conditions, 16 were classified correctly and 5 incorrectly, as shown in brackets (16/5). For exemplification of using the logistic regression method further consider that the individual in question has no dyspnoea when walking, no cardiac diseases, no waking up drowning and smoking intensity equal to 10 pack-years. Applying the equation of table 2 gives that Logit = – 5.09 + 5.15 - 1.70 = - 1.65, or equivalently, implying that the individual would be classified as Non-COPD with only a probability of 16% for COPD.

The predictive power of the five proposed methods is shown in table 4. The model validation is provided both on the total dataset and on the 10-fold cross validation. The values of sensitivity, false positive rate, F-measure, Kappa's coefficient and AUROC are shown in table 4a while the number of people classified as COPD and non-COPD based on actual values observed in the sample are shown in table 4b.

Discussion

This study shows that a simple tool consisting of symptom-based questions can be very useful in the identification of COPD patients with a smoking history in primary care. The results enable high efficiency prediction of COPD using a reduced number of factors which may be easily employed in the field of primary health care.

COPD shows high prevalence in smokers and many authors agree it is paramount to anticipate and improve diagnosis from primary care [6,8,9,21]. In our study, a percentage of 11.2% individuals affected by COPD were obtained within smoking participants, none of whom had been previously diagnosed with respiratory disease. We thus consider that a simple questionnaire could be an important tool to obtain early diagnosis from primary care and under-diagnosis reduction. Other similar studies obtain higher percentages of under-diagnosed individuals in primary health care [6,7]. However, Gingter, et al. [22] presented smaller values than ours.

There is controversy regarding the best way to face the problem of under-diagnosis and late diagnosis of COPD. The administration of questionnaires to the general population (active search), as well as to population consulting for any cause (opportunistic search) allows us to select a population with a higher risk of COPD and improves diagnostic performance of spirometry [23,24].

Using the results obtained in the 10-fold cross validation, all models have high sensitivity and AUROC. PART decision list presents

Table 1: Percentage of COPD and non-COPD in each of the population characteristics.

Item			COPD (%)	Non COPD (%)	p-value	OR	AUROC
Sex	Male		14.8	85.2	0.005	3.13	0.613
	Female		5.3	94.7			
Age Group (Years)	< 50		3.2	96.8	< 0.001	-	0.712
	50 -60		15.3	84.7			
	≥ 60		21.9	78.1			
Smoking intensity (pack-years)	< 15		1.1	98.9	< 0.001	-	0.721
	15-30		4.4	95.6			
	≥ 30		19.7	80.3			
Cough	In the morning	Yes	33.7	66.3	< 0.001	9.05	0.734
		No	5.3	94.7			
	Day	Yes	47.3	52.7	< 0.001	15.48	0.748
		No	5.5	94.5			
	Daily for three months	Yes	44.0	56.0	< 0.001	7.93	0.603
			9.0	91.0			
Dyspnoea	Walking (without other diseases)	Yes	81.0	19.0	< 0.001	53.58	0.683
		No	7.3	92.7			
	Climbing one flight of steps (at a normal rate)	Yes	68.9	31.1	< 0.001	249.05	0.940
		No	0.9	99.1			
	Walking on level ground	Yes	77.4	22.6	< 0.001	57.14	0.757
		No	5.7	94.3			
	Dressing or washing	Yes	87.5	12.5	< 0.001	65.58	0.576
		No	9.6	90.4			
	Non cardiac	Yes	10.0	90.0	0.690	0.88	0.501
			11.2	88.8			
Wheeze	Yes		35.1	64.9	< 0.001	8.81	0.722
	No		5.8	94.2			
Phlegm	In the morning	Yes	30.2	69.8	< 0.001	5.20	0.649
		No	7.7	92.3			
	Day	Yes	25.9	74.1	0.022	3.10	0.550
		No	10.1	89.9			
	Daily for three months	Yes	18.2	81.8	0.223	1.84	0.519
			10.8	89.2			
Asthma	Diagnosed	Yes	28.6	71.4	0.021	3.51	0.546
		No	10.2	89.8			
	Family history	Yes	11.8	88.2	0.476	1.08	0.507
			11.0	89.0			
Physical fitness	Oppression	Yes	55.2	44.8	< 0.001	14.60	0.66
		No	7.8	92.2			
	Spontaneous drowning	Yes	25.0	75.0	0.060	2.85	0.535
		No	10.5	89.5			
	Drowning on exertion	Yes	58.8	41.2	< 0.001	14.29	0.601
		No	9.1	90.9			
	Waking up drowning	Yes	24.0	76.0	0.048	2.74	0.540
		No	10.3	89.7			
	General physical fitness	Normal	9.6	90.4	0.024	0.407	0.569
			20.7	79.3			
Cardio-respiratory history	Rhinitis and/or Sinusitis	Yes	11.4	88.6	0.954	1.02	0.502
		No	11.1	88.9			
	Heart disease	Yes	32.0	68.0	0.003	4.32	0.565
		No	9.8	90.2			
	Pulmonary disease	Yes	37.5	62.5	0.017	5.03	0.526
		No	10.7	89.3			
	Pneumonia	Yes	28.1	71.9	0.005	3.63	0.568
			9.7	90.3			

OR: odds ratios; AUROC: Area under the receiver operating characteristic curve.

the worst values. JRip rule has the best sensitivity (0.911) and second best AUROC (0.932), only exceeded by logistic regression with an AUROC equal to 0.965. Of the 45 COPD present in the sample, 41 are correctly classified with JRip rule while 17 non-COPD will be wrongly classified as COPD. Thus, this method has the worst false positive rate (0.048). J48 decision tree has the best false positive rate (0.02) which has an influence on the possession of the best F-measure (0.818) and Kappa's coefficient (0.7958). We also found that of the 357 non-COPD only 7 are misclassified.

All methods have proved capable of discriminating individuals with or without COPD. However, considering that having COPD is the key prediction in this biomedical application, a classification method with higher sensitivity is desired. We consider JRIP rule as a discriminatory tool to order a spirometry to be the most effective. We have selected this method for having the highest sensitivity and one of the best AUROC of the methods tested once evaluated by means of a 10-fold cross validation. In addition, this model selects only five variables: dyspnoea when climbing one flight of steps, dyspnoea when walking, phlegm daily for three months, smoking intensity and cough daily for three months.

We propose a simple predictive model with a series of items easy to obtain which have proven capable to identify patients with or without COPD, so it could be used in this level of health care to

Table 2: Models obtained by applying the different data mining methods.

CART decision tree

```
IF (Dyspnoea when climbing one flight of steps = No) THEN COPD = No (338/3)
IF (Dyspnoea when climbing one flight of steps = Yes)
    IF (Age group = ( < 50))
        IF (Asthma family history= Yes) THEN COPD = No (6/0)
        IF (Asthma family history = No)
            IF (Phlegm in the morning = Yes)
                IF (Phlegm day = Yes) THEN COPD = No (3/0)
                IF (Phlegm day = No) THEN COPD = Yes (3/2)
            IF (Phlegm in the morning = No) THEN COPD = Yes (2/0)
        IF (Age group = ( ≥ 50)) THEN COPD = Yes (37/8)
```

J48 decision tree

```
IF (Dyspnoea when climbing one flight of steps = No) THEN COPD = No (341/3)
IF (Dyspnoea when climbing one flight of steps = Yes)
    IF (Age group = ( < 50)) THEN COPD = No (16/5)
    IF (Age group = (50 - 60))
        IF (Phlegm daily for three months = Yes) THEN COPD = No (4/1)
        IF (Phlegm daily for three months = No) THEN COPD = Yes (17/1)
    IF (Age group = ( > 60)) THEN COPD = Yes (24/4)
```

JRip rule

```
IF ((Dyspnoea when climbing one flight of steps = Yes) AND (Dyspnoea when walking = Yes)) THEN COPD = Yes (28/5)
IF ((Dyspnoea when climbing one flight of steps = Yes) AND (Phlegm daily for three months = No) AND (Smoking intensity = ( ≥ 30)) THEN COPD = Yes (20/4)
IF ((Cough daily for three months = Yes) AND (Phlegm daily for three months =No)) THEN COPD = Yes (2/0)
COPD = No (352/4)
```

PART decision list

```
IF (Dyspnoea when climbing one flight of steps = No) THEN COPD = No (341/3)
IF ((Heart disease = No) AND (Age group = ( ≥ 60))) THEN COPD = Yes (20/4)
IF ((Heart disease = No) AND (Age group = (50 - 60)) AND (Phlegm daily for three months = No)) THEN COPD = Yes (14/1)
IF (Heart disease = Yes) THEN COPD = Yes (8/0)
IF (Asthma family history = Yes) THEN COPD = No (8/0)
IF (Cough in the morning = Yes) THEN COPD = No (8/2)
COPD = Yes (3/0)
```

Logistic regression

Logit pCOPD = In pCOPD /(1 -pCOPD) = -5.06 + 5.15*(Dyspnoea when climbing one flight of steps = Yes) + 1.86*(Dyspnoea when walking = Yes) + 1.89*(Heart disease = Yes) -2.16*(Waking up drowning = Yes) + 1.08*(Smoking intensity = (> 30)) - 1.70* (Age group = (< 50))

pCOPD = probability of COPD, In: natural logarithm (IF Logit pCOPD > 0 THEN COPD = Yes)

Numbers in brackets, say (*a/b*), represent that there is *a* individuals of the sample with 10-fold cross validation well classified and *b* misclassified when the indicated decision is made.

Table 3: Selected variables once applied the wrapper-based approach with the respective data mining method and genetic search algorithm.

Method	Number of variables	Selected variables
CART	14	Age group, Wheeze, Dyspnoea (climbing one flight of steps), Phlegm (in the morning), Phlegm (Day), Phlegm (daily for three months), Family history of asthma, Oppression, Spontaneous drowning, Waking up drowning, General physical fitness, Rhinitis and/or Sinusitis, Heart disease, Pulmonary disease.
J48	9	Sex, Age group, Dyspnoea (climbing one flight of steps), Dyspnoea (walking on level ground), Phlegm (daily for three months), Diagnosed asthma, Spontaneous drowning, Waking up drowning, Heart disease.
JRip	12	Smoking intensity, Cough (in the morning), Cough (daily for three months), Dyspnoea (climbing one flight of steps), Dyspnoea (walking on level ground), Dyspnoea (walking without other disease), Dyspnoea (dressing or washing), Phlegm (daily for three months), Diagnosed asthma, Drowning on exertion, Pulmonary disease, Pneumonia.
PART	13	Age group, Cough (in the morning), Dyspnoea (climbing one flight of steps), Dyspnoea (dressing or washing), Phlegm (daily for three months), Diagnosed asthma, Family history of asthma, Spontaneous drowning, Waking up drowning, Rhinitis and/or Sinusitis, Heart disease, Pulmonary disease, Pneumonia.
Logistic regression	9	Age group, Dyspnoea (climbing one flight of steps), Dyspnoea (walking on level ground), Dyspnoea (non cardiac), Diagnosed asthma, Drowning on exertion, Waking up drowning, Rhinitis and/or Sinusitis, Heart disease.

detect high risk individuals where it would be indicated to perform a spirometry to allow an early diagnosis.

In order to actively search cases of COPD, Price, et al. [6] propose a questionnaire which does not comprise dyspnoea, indicating that among the six dyspnoea items, including the MRC Dyspnoea Scale, only one showed any discriminatory power (dyspnoea more in recent years). We do find significant differences in the varied aspects of presenting dyspnoea, excepting non cardiac dyspnoea (Table 1), having been selected in this model as the ones with a higher predictive power: dyspnoea when climbing one flight of steps and dyspnoea when walking.

In table 1, we can see that phlegm in the morning and during the day prevails in smokers diagnosed with COPD. However, phlegm daily for three months does not present any significant differences between COPD and non COPD groups. Note that phlegm daily for three months is a variable which was not significantly related to the presence of COPD (table 1, p = 0.223) but when analyzed in the group of those presenting dyspnoea when climbing one flight of steps it has a protective effect. Regarding the available sample, in the group with phlegm daily for three months and dyspnoea when climbing one flight of steps 33% present COPD whereas in those without phlegm daily for three months this percentage rises to 76% (p = 0.021). Price [6] indicates that of seven questions on phlegm, only two were significant, although with associations in different directions. Phlegm in the absence of a cold was strongly associated with COPD. Conversely, phlegm in the morning was not discriminatory in bivariate analysis

Table 4: Evaluation results, a) Coefficient; b) Number of people classified, of predictive models on the total dataset and after applying the 10-fold cross validation.

a)	Method	Sensitivity	1-specificity (False Positive Rate)	F-Measure	Kappa	AUROC
Use full training set	CART	0.933	0.028	0.866	0.8477	0.956
	J48	0.800	0.014	0.837	0.8178	0.956
	Jrip	0.911	0.025	0.863	0.8449	0.944
	PART	0.889	0.014	0.889	0.8749	0.962
	Logistic regression	0.822	0.017	0.841	0.8214	0.972
10-fold cross validation	CART	0.844	0.034	0.800	0.7733	0.900
	J48	0.800	0.020	0.818	0.7958	0.900
	Jrip	0.911	0.048	0.796	0.7667	0.932
	PART	0.778	0.028	0.778	0.7498	0.882
	Logistic regression	0.822	0.025	0.813	0.7893	0.965

b)	CART			J48		JRip		PART		LR	
	Classified as										
	Real	COPD	Non-COPD	COPD	Non-COPD	COPD	Non-COPD	COPD	Non-COPD	COPD	Non-COPD
Use full training set	COPD	42	3	36	9	41	4	40	5	37	8
	Non-COPD	10	347	5	352	9	348	5	352	6	351
10-fold cross validation	COPD	38	7	36	9	41	4	35	10	37	8
	Non-COPD	12	345	7	350	17	340	10	347	9	348

but showed a negative association with COPD in the final model and chronic phlegm, while strongly associated with obstruction, identified less than 1% of those with a study diagnosis of COPD.

Of the three items related to cough, cough daily for three months is the only one that has been included in the final model. Price, et al. [6] indicates that cough is the most prevalent symptom in smokers, with or without COPD, so it does not present high discriminatory power. Freeman, et al. [7] include coughing occasionally or more often as a symptom with high discriminatory power to diagnose COPD.

The intensity of tobacco consumption expressed in pack-years is another item included in the model. Most authors agree the main risk factor for this disease is the intensity of tobacco consumption [25] and it appears in the screening models proposed by different authors [6-9,26] who include age as a predicting factor. In our study, although COPD presence increases significantly with age, it does not appear as selected item in the JRip model.

Limitations of the study

The voluntary nature of the participants who joined the study may not reflect the general primary care population and the value of COPD prevalence obtained may only be a rough estimate, although we consider that it may be a good indicator if a screening takes place in this environment. Taking into account the prevalence of COPD in the population, the selection criteria for the study force to work with very large samples, which is not always possible given the high rate of failure to attend scheduled appointments.

Development of such statistical tools will require an additional study, including prospective validation of items in an appropriate clinical setting and policy recommendations on the use of these predictor factors.

Conclusion

Our data confirm the presence of a high number of smokers with respiratory symptoms who are not diagnosed with COPD.

We propose a simple predictive model, with a number of easy to obtain items that have demonstrated capacity to identify patients with and without COPD. The five tested models work acceptably, and although one cannot find a method that is always the best for the classification of different datasets and criteria, JRip decision rule has been chosen to present the best sensitivity and AUROC, as well as maintaining a low false positive rate.

Performance characteristics suggest that our questionnaire could be very useful in primary healthcare to enhance efficiency and diagnostic accuracy of current screening efforts using spirometry alone, so our model would be useful to improve the accuracy of early diagnosis of COPD in smokers with respiratory symptoms.

Authors' contributions

CRA, FRP, IGM, BCH, AAR has participated in the conception, implementation and implementation of the study, the interpretation of the results and has written the manuscript. BCH have participated in the implementation of the study, the interpretation of results and have written the manuscript. EGD has participated in the design of the study, statistical analysis, the interpretation of the results and has written the manuscript. All authors have read and approved the final manuscript.

Funding

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Ethical Approval

University of La Laguna Committee for Health Research Ethics approved the study.

Conflict of interest

None.

References

1. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, et al. (2012) Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 380: 2095-2128.

2. Vestbo J, Hurd SS, Agustí AG, Jones PW, Vogelmeier C, et al. (2013) Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. Am J Respir Crit Care Med 187: 347-365.

3. Punekar YS, Shukla A, Müllerova H (2014) COPD management costs according to the frequency of COPD exacerbations in UK primary care. Int J Chron Obstruct Pulmon Dis 9: 65-73.

4. Miravitlles M, Calle M, Soler-Cataluña JJ (2012) Clinical phenotypes of COPD: identification, definition and implications for guidelines. Arch Bronconeumol 48: 86-98.

5. Qaseem A, Wilt TJ, Weinberger SE, Hanania NA, Criner G, et al. (2011) Diagnosis and management of stable chronic obstructive pulmonary disease: a clinical practice guideline update from the American College of Physicians, American College of Chest Physicians, American Thoracic Society, and European Respiratory Society. Ann Intern Med 155: 179-191.

6. Price DB, Tinkelman DG, Halbert RJ, Nordyke RJ, Isonaka S, et al. (2006) Symptom-based questionnaire for identifying COPD in smokers. Respiration 73: 285-295.

7. Freeman D, Nordyke RJ, Isonaka S, Nonikov DV, Maroni JM, et al. (2005) Questions for COPD diagnostic screening in a primary care setting. Respir Med 99: 1311-1318.

8. Dirven JA, Tange HJ, Muris JW, van Haaren KM, Vink G, et al. (2013) Early detection of COPD in general practice: patient or practice managed?

A randomised controlled trial of two strategies in different socioeconomic environments. Prim Care Respir J 22: 331-337.

9. Dirven JA, Tange HJ, Muris JW, van Haaren KM, Vink G, et al. (2013) Early detection of COPD in general practice: implementation, workload and socioeconomic status. A mixed methods observational study. Prim Care Respir J 22: 338-343.

10. López Varela MV, Montes de Oca M, Halbert R, Muiño A, Tálamo C, et al. (2013) Comorbidities and health status in individuals with and without COPD in five Latin American cities: the PLATINO study. Arch Bronconeumol 49: 468-474.

11. Al-ani S, Spigt M, Hofset P, Melbye H (2013) Predictors of exacerbations of asthma and COPD during one year in primary care. Fam Pract 30: 621-628.

12. Kotz D, Nelemans P, van Schayck CP, Wesseling GJ (2008) External validation of a COPD diagnostic questionnaire. Eur Respir J 31: 298-303.

13. Austin PC (2007) A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. Stat Med 26: 2937-2957.

14. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, et al. (2012) Data mining in healthcare and biomedicine: a survey of the literature. J Med Syst 36: 2431-2448.

15. Tapak L, Mahjub H, Hamidi O, Poorolajal J (2013) Real-data comparison of data mining methods in prediction of diabetes in iran. Healthc Inform Res 19: 177-185.

16. Lee BJ, Kim JY (2014) A comparison of the predictive power of anthropometric indices for hypertension and hypotension risk. PLoS One 9: e84897.

17. Minette A (1989) Questionnaire of the European Community for Coal and Steel (ECSC) on respiratory symptoms. 1987--updating of the 1962 and 1967 questionnaires for studying chronic bronchitis and emphysema. Eur Respir J 2: 165-177.

18. Peces-Barba G, Barberà JA, Agustí A, Casanova C, Casas A, et al. (2008) Diagnosis and management of chronic obstructive pulmonary disease: joint guidelines of the Spanish Society of Pulmonology and Thoracic Surgery (SEPAR) and the Latin American Thoracic Society (ALAT). Arch Bronconeumol 44: 271-281.

19. Witten IH, Frank E, Hall MA (2011) Data Mining: Practical Machine Learning Tools and Techniques. (3rd edn), Morgan Kaufmann, San Francisco.

20. Chen N, Ribeiro B, Vieira AS, Duarte J, Neves JC (2011) A genetic algorithm-based approach to cost-sensitive bankruptey prediction. Expert Syst Appl 38: 12939-12945.

21. Bellamy D, Smith J (2007) Role of primary care in early diagnosis and effective management of COPD. Int J Clin Pract 61: 1380-1389.

22. Gingter C, Wilm S, Abholz HH (2009) Is COPD a rare disease? Prevalence and identification rates in smokers aged 40 years and over within general practice in Germany. Fam Pract 26: 3-9.

23. Castillo D, Guayta R, Giner J, Burgos F, Capdevila C, et al. (2009) COPD case finding by spirometry in high-risk customers of urban community pharmacies: a pilot study. Respir Med 103: 839-845.

24. Schirnhofer L, Lamprecht B, Firlei N, Kaiser B, Buist AS, et al. (2011) Using targeted spirometry to reduce non-diagnosed chronic obstructive pulmonary disease. Respiration 81: 476-482.

25. Burkhardt R, Pankow W (2014) The diagnosis of chronic obstructive pulmonary disease. Dtsch Arztebl Int 111: 834-845.

26. van Schayck CP, Halbert RJ, Nordyke RJ, Isonaka S, Maroni J, et al. (2005) Comparison of existing symptom-based questionnaires for identifying COPD in the general practice setting. Respirology 10: 323-333.