Respiratory Medicine 186 (2021) 106528

Contents lists available at ScienceDirect

Respiratory Medicine

journal homepage: www.elsevier.com/locate/rmed

Original Research

The cardiovascular phenotype of Chronic Obstructive Pulmonary Disease (COPD): Applying machine learning to the prediction of cardiovascular comorbidities

Vasilis Nikolaou [a,*], Sebastiano Massaro [a,b], Wolfgang Garn [a], Masoud Fakhimi [a], Lampros Stergioulas [c], David Price [d,e,f]

[a] University of Surrey, Surrey Business School, Guildford, GU2 7HX, United Kingdom
[b] The Organizational Neuroscience Laboratory, London, WC1N 3AX, United Kingdom
[c] The Hague University of Applied Sciences, Johanna Westerdijkplein, 75, 2521, EN Den Haag, Netherlands
[d] Optimum Patient Care, Cambridge, UK
[e] Observational and Pragmatic Research Institute, Singapore
[f] Centre of Academic Primary Care, Division of Applied Health Sciences, University of Aberdeen, Aberdeen, United Kingdom

ARTICLE INFO

Keywords:
Cardiovascular subtypes
Machine learning
Cluster analysis
Random forest

ABSTRACT

Background: Chronic Obstructive Pulmonary Disease (COPD) is a heterogeneous group of lung conditions that are challenging to diagnose and treat. As the presence of comorbidities often exacerbates this scenario, the characterization of patients with COPD and cardiovascular comorbidities may allow early intervention and improve disease management and care.

Methods: We analysed a 4-year observational cohort of 6883 UK patients who were ultimately diagnosed with COPD and at least one cardiovascular comorbidity. The cohort was extracted from the UK Royal College of General Practitioners and Surveillance Centre database. The COPD phenotypes were identified prior to diagnosis and their reproducibility was assessed following COPD diagnosis. We then developed four classifiers for predicting cardiovascular comorbidities.

Results: Three subtypes of the COPD cardiovascular phenotype were identified prior to diagnosis. Phenotype A was characterised by a higher prevalence of severe COPD, emphysema, hypertension. Phenotype B was characterised by a larger male majority, a lower prevalence of hypertension, the highest prevalence of the other cardiovascular comorbidities, and diabetes. Finally, phenotype C was characterised by universal hypertension, a higher prevalence of mild COPD and the low prevalence of COPD exacerbations. These phenotypes were reproduced after diagnosis with 92% accuracy. The random forest model was highly accurate for predicting hypertension while ruling out less prevalent comorbidities.

Conclusions: This study identified three subtypes of the COPD cardiovascular phenotype that may generalize to other populations. Among the four models tested, the random forest classifier was the most accurate at predicting cardiovascular comorbidities in COPD patients with the cardiovascular phenotype.

1. Introduction

Chronic Obstructive Pulmonary Disease (COPD) comprises a group of lung diseases, including asthma, emphysema and chronic bronchitis, that cause breathing difficulties due to inflammation of the lungs and narrowing of the airways [1]. According to the World Health Organisation (WHO), COPD is projected to become the third leading cause of death by 2030 [2] because our ability to diagnose early and treat effectively has been relatively static. To better understand the heterogeneity of COPD, recent and ongoing research [3] is applying a wide range of machine learning methods, which can integrate patients' demographic and clinical characteristics to derive underlying disease traits that often occur together (i.e., COPD phenotypes). Among these, the cardiovascular phenotype remains one of the most relevant phenotypes to analyse, given that cardiovascular disease is the major contributor to morbidity and mortality in patients with COPD [4]. Unfortunately, however, this phenotype is highly complex and variegated being

* Corresponding author. University of Surrey, Surrey Business School, Alexander Fleming Rd, Guildford, GU2 7XH, United Kingdom.
E-mail address: v.nikolaou@surrey.ac.uk (V. Nikolaou).

https://doi.org/10.1016/j.rmed.2021.106528
Received 28 April 2021; Received in revised form 29 June 2021; Accepted 1 July 2021
Available online 7 July 2021
0954-6111/© 2021 Elsevier Ltd. All rights reserved.

Abbreviations

COPD	Chronic Obstructive Pulmonary Disease
FEV1	Forced Expiratory Volume in 1 s
FVC	Forced Vital Capacity
GP	General Practitioner
ICS	Inhaled Corticosteroids
LABA	Long-Acting Beta Agonist
LAMA	Long-Acting Anti-Muscarinic
MCA	Multiple Correspondence Analysis
MICE	Multivariate Imputation by Chained Equations
NPV	Negative Predictive Value
PPV	Positive Predictive Value
RF	Random Forest
RCGP	Royal College of General Practitioners
RSC	Research and Surveillance Centre
SAMA	Short-Acting Anti-Muscarinic
WHO	World Health Organisation

characterised by substantial differences in age, sex, and the hospital admission rate for acute exacerbations of COPD [5–7]. It thus remains both paramount and challenging to predict which COPD patients will develop cardiovascular comorbidities in the future.

This study aims to address this gap by characterising subtypes of the COPD cardiovascular phenotype. We derive three subtypes from a cohort of patients diagnosed with cardiovascular comorbidities before COPD and reproduce the subtypes in a cohort of patients after COPD diagnosis. Then, we train and test four classifiers to optimise the prediction of cardiovascular comorbidities in COPD patients.

2. Methods

2.1. Study design

This is a retrospective analysis of an observational cohort of patients with COPD in the UK. The data covers a 4-year period (2015–2018) and was extracted from the Royal College of General Practitioners (RCGP) Research and Surveillance Centre (RSC) database [8,9], which includes more than 5 million patients, over 2 million records, and 500 million prescriptions (as of December 2017) [10]. This project was approved by the University of Surrey's Institutional Review Board (353,003-352, 994-40371,074).

2.2. Study population

Fig. 1 shows the inclusion and exclusion criteria, which yielded 6883 patients.

To be included, a patient needed to have a Read code [11] for COPD diagnosis, a diagnosis of at least one cardiovascular comorbidity, be older than 35 years of age, be a current or former smoker (i.e., ex-smoker), not have active asthma, have a Forced Expiratory Volume in 1 s to Forced Vital Capacity Ratio (FEV1/FVC ratio) of less than or equal to 0.7 (i.e., the threshold for COPD diagnosis [1]) and have follow-up FEV1 values recorded for 3 consecutive years. Recent research confirms that a period of 3 years is an ideal timespan to account for clinically relevant FEV1 variations in COPD patients [12].

We excluded patients who met one of the following: less than 35 years of age, never-smoker, active asthma, FEV1/FVC ratio greater than 0.7 and lacking 3 consecutive years of FEV1 tests.

2.3. Statistical analysis

We split our sample into two cohorts: a) the training cohort,

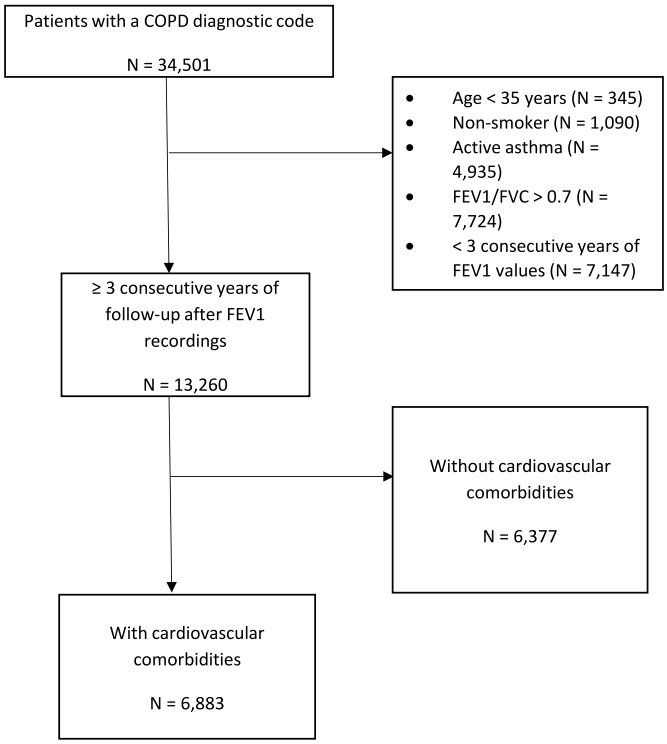

Fig. 1. Flow chart of the study cohort.

consisting of patients who were registered with a GP before the COPD diagnosis, and b) the validation cohort, consisting of patients who were not registered until after their COPD diagnosis (Fig. 2). Splitting the sample into two independent cohorts on the basis of such a clear-cut objective criterion (i.e., before and after COPD diagnosis), rather than randomly, allows the algorithms to unambiguously learn how to identify COPD phenotypes and classify patients into cardiovascular comorbidities at an early stage of the disease. In other terms, this is because the algorithms' learning step occurs among patients not yet diagnosed with COPD. We then used the training clusters (i.e., those clusters learned prior to diagnosis) to predict new clusters in the cohort of patients after COPD diagnosis, and assessed their agreement as described below in the "Cluster validation" section. Similarly, we used the classification of patients into four cardiovascular comorbidities learned by the algorithms in the training cohort to predict new classes of cardiovascular comorbities in the validation cohort. Finally, we assessed the validity of the predicted classes by cross-examining them with the pre-existing (i.e., observed) cardiovascular comorbidities.

To perform these analyses, we used two types of machine learning approaches well suited to: a) identify clusters (i.e., subtypes) of the cardiovascular phenotype, and b) predict cardiovascular comorbidities in a new cohort of patients with COPD. For the first objective, we used unsupervised learning where we had no prior knowledge of the classification of patients into clusters. Indeed, these clusters are just inferred from the relationships within the data, and they are the algorithms which assign labels to the derived phenotypes (see the "Clustering" section below). To predict cardiovascular comorbidities, our second goal, we instead used supervised learning. Here, the classification of patients into cardiovascular comorbidities was already known a priori from the dataset, and our aim was to predict future classes (i.e., cardiovascular comorbidities) in a new (blind) cohort (i.e., the cohort after COPD diagnosis). The classification algorithms that we used for this task are further described in the "Predictive models" section of this paper.

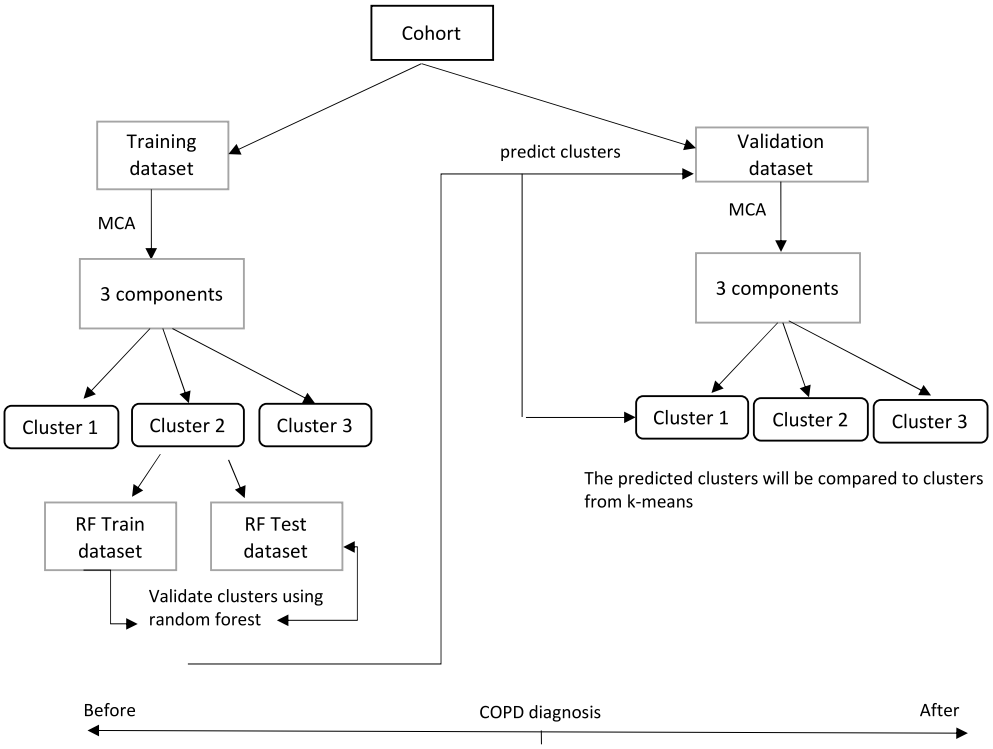

Fig. 2. Main steps in phenotype identification before and after COPD diagnosis.

2.4. Data reduction

We used multiple correspondence analysis (MCA) [13] to reduce the dimensionality of the training cohort from 19 variables (sex, body mass index, smoking, COPD severity, COPD exacerbations, emphysema, diabetes, hypertension, coronary artery disease, acute myocardial infarction, congestive cardiac failure, anxiety, depression and six types of treatment) into three uncorrelated components. We then applied k-means cluster analysis to the three components to identify the groups of patients with similar characteristics (i.e., subtypes of the COPD cardiovascular phenotype). We imputed missing values for body mass index and COPD severity with Multivariate Imputation by Chained Equations (MICE) [14].

2.5. Clustering

We used a hierarchical cluster analysis [15] to visually inspect—with a dendrogram—the optimal number of clusters (Fig. 3). We then confirmed the number of clusters by performing the elbow [16] and silhouette [17] methods (Fig. 4).

Fig. 5 compares the silhouette plots of the clusters derived from two clustering methods: hierarchical (top plot) and k-means (bottom plot). Specifically, we compared a) the magnitude of the average silhouette width, and b) the sign (positive or negative) of the silhouette width. The average silhouette width was larger under the k-means algorithm than under the hierarchical algorithm. More subjects had a negative silhouette width under the hierarchical algorithm than under k-means clustering, especially for clusters 1 and 3. We concluded that k-means clustering generates more stable clusters than the hierarchical approach.

2.6. Cluster validation

After establishing the three phenotype subtypes with k-means clustering, we developed our predictive model. The Random Forest (RF) [18] model uses as independent variables (or predictors) the 19 categorical variables described above in the MCA step, with the addition of age and lung function (FEV1). First, we used what we called "the RF training dataset" (i.e., 70% of the full training dataset, randomly selected; n = 4166), to train the RF model on the clusters identified by k-means clustering [16]. Then, we tested the RF model on an holdout group of the training dataset, the "RF test dataset" (i.e., the remaining 30% of the training dataset; n = 1785) and achieved 99% accuracy.

Next, we trained the same model on the full training dataset (i.e., the RF training and test datasets combined, which is ultimately the training cohort pre-COPD diagnosis) and checked the predicted cluster assignments against the entire validation dataset, which is the cohort of patients post- COPD diagnosis (whose clusters were also derived with k-means clustering). We used the Adjusted Rand index [19] and Jaccard

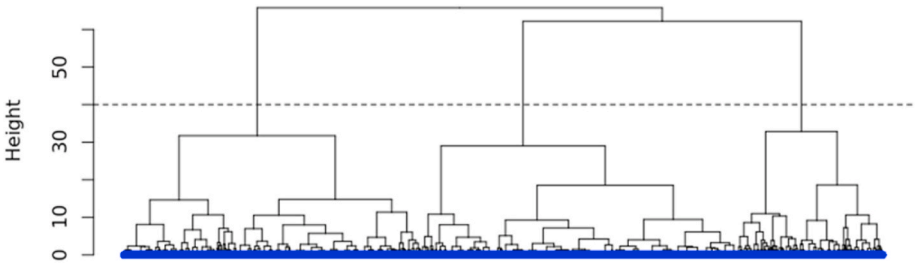

Fig. 3. Inspecting the number of clusters using hierarchical analysis in the training dataset.

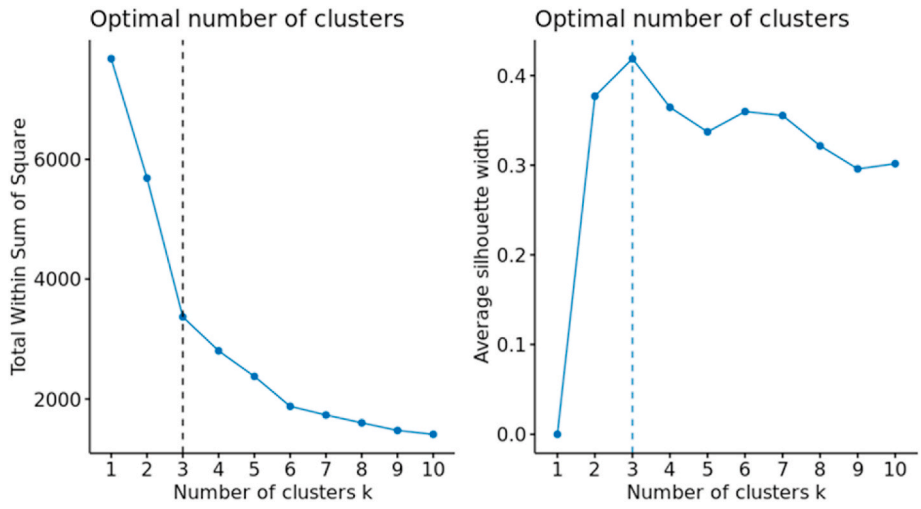

Fig. 4. Determining the optimal number of clusters for the training dataset.

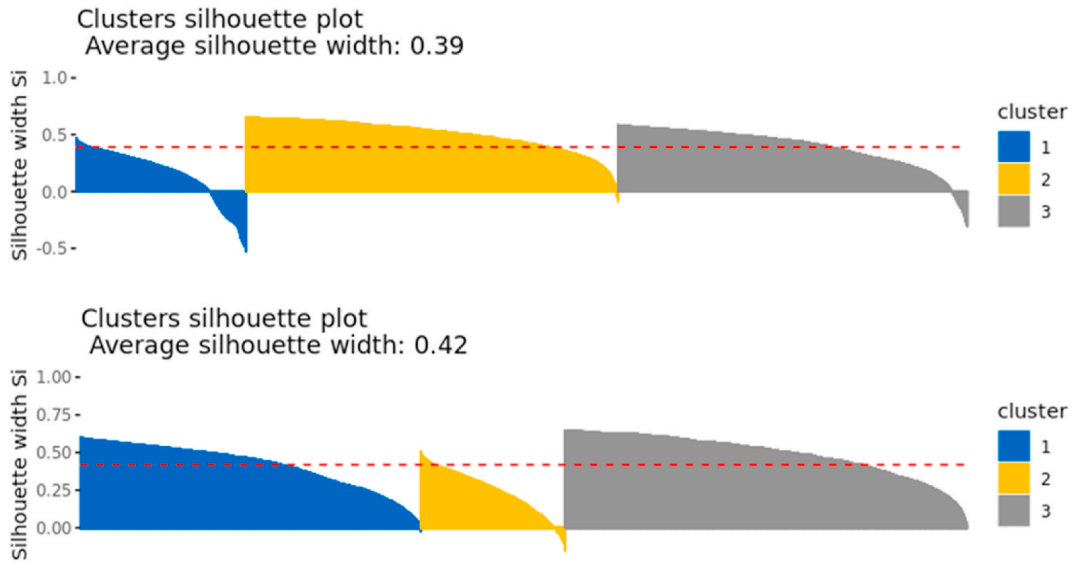

Fig. 5. Silhouette plots to determine the optimal clustering method - hierarchical (top) and k-means (bottom).

index [20] to compare the clusters predicted by the RF model with those derived by k-means clustering, and we found 92% agreement.

2.7. Predictive models

With three highly robust COPD cardiovascular phenotype subtypes established, we proceeded to train four different classifiers to predict cardiovascular comorbidities from other components of the phenotype (i.e., demographics, COPD severity, and COPD treatments). Specifically, we trained a decision tree, multinomial logistic regression, RF and gradient boosting machine [21]. We were interested in predicting four cardiovascular comorbidities: hypertension, coronary artery disease, acute myocardial infarction and congestive cardiac failure. We trained each classifier on the RF training dataset and tested the optimised classifier on the RF test dataset. Once each model was finely tuned by using automated tuning within the R [22] library 'caret', we trained it on the whole training dataset and assessed its performance on the validation dataset.

All four models used cardiovascular comorbidities as the dependent variable and the following variables as predictors: age, sex, body mass index, smoking, COPD severity, COPD exacerbations, emphysema, lung function (FEV1), diabetes, anxiety, depression and type(s) of treatment (Inhaled Corticosteroids (ICS), ICS and Long-Acting Beta Agonist (LABA), Long-Acting Anti-Muscarinic (LAMA), LABA, Short-Acting Anti-Muscarinic (SAMA), mucolytics).

Moreover, in light of the class imbalance (i.e., a disparity in the distribution of patients with cardiovascular comorbidities), we re-trained the models with two sub-sampling methods: a) up-sampling, in which we randomly sampled (with replacement) the minority class until it was the same size as the majority class, and b) down-sampling, in which we randomly sampled (with replacement) the majority class until it was the same size as the minority class. The models were then evaluated on the blind validation dataset. All statistical analyses were implemented with the statistical software R [22].

3. Results

3.1. Patient characteristics

Table 1 summarizes the descriptive baseline characteristics (Year 1) of patients who were registered with a GP before their COPD diagnosis and after diagnosis.

Table 1
Baseline (Year 1) demographic and clinical characteristics of patients with cardiovascular comorbidities who established care with a GP before and after COPD diagnosis.

Variables	Prior to COPD diagnosis (n = 5951)	After COPD diagnosis (n = 932)	Total (n = 6883)
Age, mean (SD), years	72 (9)	72 (9)	72 (9)
Sex, Male, No. (%)	3580 (60)	552 (59)	4132 (60)
Body mass index, mean (SD), kg/m^2	28 (6)	27 (6)	28 (6)
Body mass index, No. (%) with data	5937 (99)	925 (99)	6862 (99)
Underweight	134 (2)	32 (3)	166 (2)
Normal weight	1719 (29)	296 (32)	2015 (29)
Overweight	2220 (37)	315 (34)	2535 (37)
Obese	1864 (31)	282 (30)	2146 (31)
Smoking status, No. (%)			
Active smoker	1884 (32)	289 (31)	2173 (32)
Former smoker	4067 (68)	643 (69)	4710 (68)
COPD severity, No. (%) with data	3064 (51)	925 (52)	3552 (52)
Mild	1012 (33)	157 (32)	1169 (33)
Moderate	1532 (50)	244 (50)	1776 (50)
Severe	477 (16)	82 (17)	559 (16)
Very severe	43 (1)	5 (1)	48 (1)
COPD exacerbations in the past year, mean (SD)	0.3 (0.9)	0.5 (1.0)	0.3 (1.0)
COPD exacerbations in the past year, No. (%)			
0	5065 (85)	728 (78)	5793 (84)
1	509 (9)	104 (11)	613 (9)
2	225 (4)	40 (4)	265 (4)
>2	152 (3)	60 (6)	212 (3)
FEV1, mean (SD), L	0.7 (0.2)	0.7 (0.2)	0.7 (0.2)
Emphysema, No. (%)	320 (5)	106 (11)	426 (6)
Diabetes, No. (%)	1322 (22)	208 (22)	1530 (22)
Hypertension, No. (%)	5317 (89)	823 (88)	6140 (89)
Coronary artery disease, No. (%)	675 (11)	106 (11)	781 (11)
Acute myocardial infarction, No. (%)	822 (14)	144 (15)	966 (14)
Congestive cardiac failure, No. (%)	719 (12)	110 (12)	829 (12)
Anxiety, No. (%)	460 (8)	76 (8)	536 (8)
Depression, No. (%)	1668 (28)	289 (31)	1957 (28)
Treatment, No. (%)[a]			
ICS	2675 (45)	527 (57)	3202 (47)
ICS + LABA	2341 (39)	481 (52)	2822 (41)
LAMA	2805 (47)	494 (53)	3299 (48)
LABA	574 (10)	85 (9)	659 (10)
SAMA	335 (6)	54 (6)	389 (6)
Mucolytics	575 (10)	114 (12)	689 (10)

ICS: Inhaled Corticosteroids; LABA: Long-Acting Beta Agonist; LAMA: Long-Acting Anti-Muscarinic; SAMA: Short-Acting Anti-Muscarinic.

3.2. Prior to COPD diagnosis

Table 2 presents the baseline characteristics of the three subtypes of the COPD phenotype among patients with cardiovascular comorbidities who established care with a GP before their COPD diagnosis.

Phenotype A was characterised by the highest prevalence of severe COPD (as defined by the physician), substantial emphysema and nearly universal hypertension (though this was also true of phenotype C). Phenotype A was the most heavily medicated; almost all patients with this phenotype were treated with ICS and/or a combination of ICS and LABA; more than half were also treated with LAMA. Phenotype B was characterised by a large majority of male patients (whereas males comprised a small majority of the other phenotypes). Phenotype B had the lowest prevalence of hypertension but the highest prevalence of coronary artery disease, acute myocardial infarction, congestive cardiac failure, and diabetes. Just under half of the phenotype B patients were treated with LAMA; the next most common medications were ICS, followed by ICS with LABA. Phenotype C was characterised by universal hypertension (similar to phenotype A), though phenotype C had the lowest prevalence of severe COPD, the highest prevalence of mild COPD and the largest majority of patients with no exacerbations in the past year. Overall, patients with phenotype C were less medicated than the other phenotypes; the most common treatment was LAMA, though only about one-third of phenotype C patients used it. The most notable characteristics of each of the three phenotypes are summarized in Table 3.

Table 2
Baseline (Year 1) phenotype characteristics prior to COPD diagnosis in patients with cardiovascular comorbidities.

Variables	Phenotype		
	A (n = 2072)	B (n = 943)	C (n = 2936)
Age, mean (SD), years	72 (8)	72 (9)	72 (9)
Sex, Male, No. (%)	1199 (58)	732 (78)	1649 (56)
Body mass index, mean (SD), kg/m^2	28 (6)	28 (5)	28 (6)
Body mass index, No. (%) with data	2067 (99)	940 (100)	2930 (99)
Underweight	56 (3)	16 (2)	62 (2)
Normal weight	595 (29)	269 (29)	855 (29)
Overweight	772 (37)	383 (41)	1065 (36)
Obese	644 (31)	272 (29)	948 (32)
Smoking status, No. (%)			
Active smoker	586 (28)	297 (31)	1001 (34)
Former smoker	1486 (72)	646 (69)	1935 (66)
COPD severity, No. (%) with data	1196 (58)	493 (52)	1375 (47)
Mild	288 (24)	154 (31)	570 (41)
Moderate	583 (49)	262 (53)	687 (50)
Severe	295 (25)	72 (15)	110 (8)
Very severe	30 (3)	5 (1)	8 (1)
COPD exacerbations in the past year, mean (SD)	0.5 (1)	0.2 (0.7)	0.1 (0.5)
COPD exacerbations in the past year, No. (%)			
0	1584 (76)	815 (86)	2666 (91)
1	239 (12)	77 (8)	193 (7)
2	128 (6)	33 (3)	64 (2)
>2	121 (6)	18 (2)	13 (1)
FEV1, mean (SD), L	0.7 (0.2)	0.7 (0.2)	0.8 (0.2)
Emphysema, No. (%)	145 (7)	54 (6)	121 (4)
Diabetes, No. (%)	444 (21)	249 (26)	629 (21)
Hypertension, No. (%)	2055 (99)	326 (35)	2936 (100)
Coronary artery disease, No. (%)	59 (3)	500 (53)	116 (4)
Acute myocardial infarction, No. (%)	93 (4)	617 (65)	112 (4)
Congestive cardiac failure, No. (%)	174 (8)	379 (40)	166 (6)
Anxiety, No. (%)	165 (8)	67 (7)	228 (8)
Depression, No. (%)	584 (28)	278 (29)	806 (27)
Treatment, No. (%)[a]			
ICS	2054 (99)	402 (43)	219 (7)
ICS + LABA	1981 (96)	353 (37)	7 (0.2)
LAMA	1451 (70)	437 (46)	917 (31)
LABA	102 (5)	81 (9)	391 (13)
SAMA	114 (6)	50 (5)	171 (6)
Mucolytics	380 (18)	104 (11)	91 (3)

ICS: Inhaled Corticosteroids; LABA: Long-Acting Beta Agonist; LAMA: Long-Acting Anti-Muscarinic; SAMA: Short-Acting Anti-Muscarinic.

3.3. Predicting cardiovascular comorbidities after COPD diagnosis

We tested the four trained classifiers on the validation dataset (i.e.,

Table 3
Phenotype characteristics of patients with cardiovascular comorbidities prior to COPD diagnosis.

Phenotype A	Phenotype B	Phenotype C
Highest prevalence of severe COPD	Larger majority of males	Lowest prevalence of severe COPD
Emphysema (more prevalent)	Highest prevalence of three cardiovascular comorbidities:	Zero COPD exacerbations (large majority)
Hypertension (almost all)	Coronary artery disease	Hypertension (all)
Most-treated overall	Acute myocardial infarction	Least-treated overall
ICS (nearly all)	Congestive cardiac failure	LAMA (one-third)
ICS + LABA (nearly all)	Highest prevalence of diabetes	
LAMA (large majority)	Intermediate level of treatment:	
Mucolytics	ICS (almost half)	
	ICS + LABA (one-third)	
	LAMA (almost half)	

ICS: Inhaled Corticosteroids; LABA: Long-Acting Beta Agonist; LAMA: Long-Acting Anti-Muscarinic; SAMA: Short-Acting Anti-Muscarinic.

post-COPD diagnosis), and we present the results in confusion matrices (Table 4). For each predictive model (i.e., each classifier), Table 4 compares the number of patients predicted to have each cardiovascular comorbidity with the actual number of diagnoses; it also reports the classifier's overall accuracy, sensitivity (i.e., the percentage of positive cases that were predicted to be positive), specificity (i.e., the percentage of negative cases that were predicted to be negative), positive predictive value (PPV, i.e., the percentage of positive predictions that were actually positive cases) and negative predictive value (NPV, i.e., the percentage of negative predictions that were actually negative cases).

As shown in Table 4, the RF classifier (even without sub-sampling) outperformed the other models. All models exhibited relatively high sensitivity and low specificity for hypertension, but the RF classifier had the highest sensitivity (87%) and PPV (98%, versus 34%–40% in the other models). All models exhibited relatively low sensitivity and high specificity for the other three cardiovascular comorbidities (coronary artery disease, acute myocardial infarction and congestive cardiac failure), but RF was the most accurate at ruling out these conditions (NPV: 99% for all three conditions, versus 74–85% in the other models).

4. Discussion

This study presents the use of machine learning toward acquiring a better characterization of the cardiovascular phenotype in patients with COPD and predicting specific cardiovascular comorbidities linked to these patients. Given the substantial contribution of cardiovascular disease to morbidity and mortality in COPD and the complexity of the cardiovascular phenotype we believe that our findings can offer several beneficial avenues to respiratory researchers and clinicians alike. For one example, by identifying subtypes of the cardiovascular phenotype and predicting future cardiovascular comorbidities early (i.e. prior to COPD diagnosis), it is possible to better understand of the disease's development, and consequently improve disease management, possibly prevent the development of cardiovascular disease, and thus lead to the application as well as development of targeted treatments.

Here, we specifically examined four cardiovascular comorbidities—hypertension, coronary artery disease, acute myocardial infarction and congestive cardiac failure—and used basic demographic information, COPD severity, and types of COPD treatments to predict a patient's phenotype. Two of the phenotypes (A and C) had almost universal hypertension but differed in COPD severity and treatment. Meanwhile, the third phenotype (B) had a lower prevalence of hypertension but a higher prevalence of coronary artery disease, acute myocardial infarction and congestive cardiac failure, as well as diabetes.

The large size of our training sample enabled the model to predict patients' phenotypes with high accuracy (92%). This encouraging result suggests that the three identified phenotypes may generalize to other datasets and populations of patients with COPD. Our use of statistical and machine learning tools went beyond a traditional summary of the demographic and clinical characteristics of patients with COPD, which offer little in the way of predictive diagnostics. We tested several algorithms, from a conventional multinomial logistic regression model to stronger classifiers such as the RF and gradient boosting machine, which are ensembles of weaker classifiers (i.e., classifiers with low predictive power such as decision trees are combined into classifiers with stronger predictive ability).

Moreover, we handled incomplete observations with multiple imputation, and we addressed class imbalance (i.e., unequal numbers of patients with each cardiovascular comorbidity) with additional sampling methods (namely, up- and down-sampling). We assessed the performance of our four candidate models by calculating the overall accuracy (86% for RF) as well as the sensitivity, specificity, PPV, and NPV for each comorbidity. The data showed that all four classifiers, and RF in particular, were highly sensitive in predicting hypertension (highly prevalent in phenotypes A and C) and highly specific in predicting the other three (less prevalent) cardiovascular comorbidities (coronary artery disease, acute myocardial infarction and congestive cardiac failure). These findings are of substantial clinical importance because these algorithms can be used as diagnostic tools for preventing cardiovascular disease. We indeed note that the information inputed in the models is readily acquirable during any medical visit, hence offering the opportunity of rapid implementation of our framework in the clinical practice toward anticipatory diagnosis and improved medical predictions.

Finally, our findings suggest that patients clustered into three cardiovascular phenotypes also had different treatment patterns. Specifically, patients with less severe COPD (phenotype C) received less treatments; those with high prevalence of coronary artery disease, acute myocardial infarction and congestive cardiac failure and diabetes had an intermediate level of treatment (phenotype B); and, those with more severe COPD were the most-treated (phenotype A). These results are also clinically salient because they can assist clinicians to differentially treat these groups of patients, thus minimizing costs and adverse events of less-effective treatments. This categorization will also help future research toward the development of personalized therapies based on the patients' phenotype characteristics.

4.1. Limitations

We acknowledge four main limitations of this work that however represent important calls for future research. First, cluster analysis is a data-driven machine learning method; for this reason, the clusters (i.e., the phenotypes) derived bring no substansive meaning. They are formed by identifying groups of patients with similar characteristics (i.e., phenotype A, B or C); however the clinician still has to meaningfully interpret and label those clusters. While this interpretation remains a subjective task within the medical encounter, our categorization here provides a blueprint toward a more refined and standardized understanding of the heterogenous nature of the disease. Future research is thus tasked to provide clinical consensus to the meaning of the phenotypes identified in this work to enable their implementations in the everyday medical practice. Second, we considered patients with at least three consecutive years of follow-up spirometry data because this allowed us to assess more reliable lung function measures and feed more complete lung function data into the predictive models. Including patients with different follow-up times - which often happens in real clinical practice - could have given us different results. Future research may test the robustness of our results by performing a sensitivity analysis by including those patients with less follow-up period of lung function recordings. Third, the RCGP database lacked data on relevant

Table 4
Confusion matrices of four models predicting cardiovascular comorbidities in patients with COPD.

Random Forest (no sampling)		Observed			
		Hypertension	Coronary artery disease	Acute myocardial infraction	Congestive cardiac failure
Predicted	**Hypertension**	3382	19	12	21
	Coronary artery disease	156	4	0	0
	Acute myocardial infraction	188	0	4	1
	Congestive cardiac failure	157	0	2	0
Statistics	**Accuracy (%) (95% CI)**	86 (85, 87)			
	Sensitivity (%)	87	17	22	0
	Specificity (%)	17	96	95	96
	PPV (%)	98	3	2	0
	NPV (%)	2	99	99	99
Decision Tree (up-sampling)		**Hypertension**	**Coronary artery disease**	**Acute myocardial infraction**	**Congestive cardiac failure**
Predicted	**Hypertension**	1193	752	738	751
	Coronary artery disease	64	53	19	24
	Acute myocardial infraction	53	47	40	53
	Congestive cardiac failure	42	34	34	49
Statistics	**Accuracy (%) (95% CI)**	34 (32, 35)			
	Sensitivity (%)	88	6	5	6
	Specificity (%)	14	97	95	96
	PPV (%)	35	33	21	31
	NPV (%)	69	78	79	78
Gradient boosting machine (up-sampling)		**Hypertension**	**Coronary artery disease**	**Acute myocardial infraction**	**Congestive cardiac failure**
Predicted	**Hypertension**	1367	895	549	623
	Coronary artery disease	46	66	21	29
	Acute myocardial infraction	57	49	42	45
	Congestive cardiac failure	51	40	20	48
Statistics	**Accuracy (%) (95% CI)**	39 (34, 40)			
	Sensitivity (%)	89	6	7	6
	Specificity (%)	15	97	95	96
	PPV (%)	40	40	22	30
	NPV (%)	70	74	84	82
Multinomial logistic regression (up-sampling)		**Hypertension**	**Coronary artery disease**	**Acute myocardial infraction**	**Congestive cardiac failure**
Predicted	**Hypertension**	1167	874	484	909
	Coronary artery disease	45	67	21	27
	Acute myocardial infraction	46	55	29	63
	Congestive cardiac failure	36	49	20	54
Statistics	**Accuracy (%) (95% CI)**	33 (32, 35)			
	Sensitivity (%)	90	6	5	5
	Specificity (%)	15	97	95	96
	PPV (%)	34	42	15	34
	NPV (%)	75	74	86	74

CI: Confidence Interval; PPV: Positive Predictive Value; NPV: Negative Predictive Value.

biomarkers, such as cytokines, that are well-known to be associated with coronary artery disease and myocardial infarction [23]. Should such biomarkers be available, our models would become even more accurate in predicting those less prevalent cardiovascular comorbidities and subsequently improve the sensitivity and PPV rates. Finally, the RCGP database covers a limited number of cardiovascular comorbidities, so the predictions are not exhaustive. All of these limitations could be addressed in the future by applying our models to other COPD datasets (e.g., the OPCRD database [24]).

5. Conclusions

To the best of our knowledge, this study is the first to implement machine learning to identify clinically meaningful phenotypes of cardiovascular comorbidities that develop after a COPD diagnosis, though we are not the first to apply machine learning to COPD in general [3].

We used k-means clustering to identify three phenotypes prior to COPD diagnosis, and we trained an RF model to predict these phenotypes in a different blind dataset (i.e., after COPD diagnosis). We achieved a high level of agreement (92%) between the predicted cluster assignments and those derived by k-means clustering. Moreover, we trained and validated four different classifiers (of which RF performed the best) to predict cardiovascular comorbidities based on patients' demographics, COPD severity, and COPD treatments. This model represents a robust preliminary framework for predicting cardiovascular comorbidities in patients with a COPD diagnosis, though the model's predictive power likely could be improved with the inclusion of other

risk factors such as biomarkers.

The insights presented in this paper may inform GPs' medical decision making for acute complaints (namely, acute myocardial infarction and congestive cardiac failure) as well as screening and prevention (for hypertension, coronary artery disease, and diabetes) in patients with a COPD diagnosis. Validation of our framework in non-UK populations may contribute to a more nuanced understanding of the COPD cardiovascular phenotypes, ultimately improving treatment for cardiovascular comorbidities in COPD patients and enabling their prevention at an earlier stage.

Guarantor statement

Vasilis Nikolaou agrees to be accountable for all content and aspects of the work, ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Author contributions

Vasilis Nikolaou had full access to and analysis of the data. All authors were involved in the conception and design of the study, the interpretation, as well the critical revision of the manuscript. Vasilis Nikolaou and Sebastiano Massaro were responsible for drafting the manuscript. The study was supervised by Wolfgang Garn, Masoud Fakhimi and Lampros Stergioulas. All authors approved the final version of this manuscript and agree to be accountable for all aspects of the work.

Funding information

This research did not receive any specific grants from funding agencies in the public, commercial or not-for-profit sectors.

Notation of prior abstract publication/presentation

None.

Clinical trial registration

None.

CRediT authorship contribution statement

Vasilis Nikolaou: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Sebastiano Massaro:** Writing – review & editing, Writing – original draft. **Wolfgang Garn:** Resources, Software, Supervision, Validation, Writing – review & editing. **Masoud Fakhimi:** Resources, Software, Supervision, Validation, Writing – review & editing. **Lampros Stergioulas:** Resources, Software, Supervision, Validation, Writing – review & editing. **David Price:** Conceptualization, Writing – review & editing.

Acknowledgments

We would like to thank patients for allowing their data to be used for surveillance and research, General Practitioners who agreed to be part of the RCGP RSC and allowed us to extract and use health data for surveillance and research, Ms. Filipa Ferreira from RCGP, Mr. Julian Sherlock from the University of Surrey, Apollo Medical Systems for data extraction, collaborators with EMIS, TPP, In-Practice and Micro-Test CMR suppliers for facilitating data extraction, and colleagues at Public Health England.

References

[1] NHS inform on Chronic obstructive pulmonary disease. https://www.nhsinform.scot/illnesses-and-conditions/lungs-and-airways/copd/chronic-obstructive-pulmonary-disease#about-copd. (Accessed 15 February 2020).
[2] World Health Organization on chronic respiratory diseases and COPD. https://www.who.int/respiratory/copd/en/. (Accessed 15 February 2020).
[3] V. Nikolaou, S. Massaro, M. Fakhimi, L. Stergioulas, D. Price, COPD phenotypes and machine learning cluster analysis: a systematic review and future research agenda, Respir. Med. (2020 Jul 28) 106093.
[4] H. Müllerova, A. Agusti, S. Erqou, D.W. Mapel, Cardiovascular comorbidity in COPD: systematic literature review, Chest 144 (4) (2013) 1163–1178.
[5] M. Pikoula, J.K. Quint, F. Nissen, H. Hemingway, L. Smeeth, S. Denaxas, Identifying clinically important COPD sub-types using data-driven approaches in primary care population based electronic health records, BMC Med. Inf. Decis. Making 19 (1) (2019 Dec) 86.
[6] P.R. Burgel, J.L. Paillasseur, B. Peene, et al., Two distinct chronic obstructive pulmonary disease (COPD) phenotypes are associated with high risk of mortality, PloS One 7 (12) (2012).
[7] A. Agusti, P.M. Calverley, B. Celli, et al., Characterisation of COPD heterogeneity in the ECLIPSE cohort, Respir. Res. 11 (1) (2010 Dec 1) 122.
[8] Royal College of general Practitioners (RCG) research and surveillance Centre (RSC):. http://www.rcgp.org.uk/rsc.
[9] S. de Lusignan, A. Correa, G.E. Smith, I. Yonova, R. Pebody, F. Ferreira, A.J. Elliot, D. Fleming, RCGP Research and Surveillance Centre: 50 years' surveillance of influenza, infections, and respiratory conditions, Br. J. Gen. Pract. 67 (663) (2017 Oct) 440–441, https://doi.org/10.3399/bjgp17X692645.
[10] A. Correa, W. Hinton, A. McGovern, J. van Vlymen, I. Yonova, S. Jones, S. de Lusignan, Royal College of general Practitioners research and surveillance Centre (RCGP RSC) sentinel network: a cohort profile, BMJ Open 6 (4) (2016 Apr 20), e011092, https://doi.org/10.1136/bmjopen-2016-011092.
[11] Coded thesaurus of clinical terms. https://digital.nhs.uk/services/terminology-and-classifications/read-codes, 2018-04-01.
[12] J. Koskela, M. Katajisto, A. Kallio, M. Kilpeläinen, A. Lindqvist, T. Laitinen, Individual FEV1 trajectories can be identified from a COPD cohort, COPD 13 (4) (2016) 425–430.
[13] Mori Y, Kuroda M, Makino N. Nonlinear principal component analysis. In Nonlinear Principal Component Analysis and its Applications 2016 (pp. 7-20). Springer, Singapore.
[14] S. van Buuren, K. Groothuis-Oudshoorn, Mice: multivariate imputation by chained Equations in R, J. Stat. Software 45 (3) (2011) 1–67. https://www.jstatsoft.org/v45/i03/.
[15] F. Murtagh, P. Legendre, Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? J. Classif. 31 (3) (2014 Oct 1) 274–295.
[16] P. Bholowalia, A. Kumar, EBK-means: a clustering technique based on elbow method and k-means in WSN, Int. J. Comput. Appl. 105 (9) (2014 Jan 1).
[17] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987 Nov 1) 53–65.
[18] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001 Oct 1) 5–32.
[19] D. Steinley, Properties of the hubert-arable adjusted Rand index, Psychol. Methods 9 (3) (2004 Sep) 386.
[20] S. Fletcher, M.Z. Islam, Comparing sets of patterns with the Jaccard index, Australasian Journal of Information Systems (2018 Mar 7) 22.
[21] J.H. Friedman, Stochastic Gradient Boosting, Comput. Stat. Data Anal. 38 (4) (2002) 367–378.
[22] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2013. http://www.R-project.org/.
[23] L. Stoner, A.A. Lucero, B.R. Palmer, L.M. Jones, J.M. Young, J. Faulkner, Inflammatory bio markers for predicting cardiovascular disease, Clin. Biochem. 46 (15) (2013 Oct 1) 1353–1371.
[24] Clinical Practice Research Datalink (CPRD, Optimum patient care research database (OPCRD;. http://www.cprd.com/. https://opcrd.co.uk/.