scientific reports

OPEN

A machine learning framework for short-term prediction of chronic obstructive pulmonary disease exacerbations using personal air quality monitors and lifestyle data

M. Atzeni[1], G. Cappon[1], J. K. Quint[3], F. Kelly[2], B. Barratt[2] & M. Vettoretti[1]✉

Chronic Obstructive Pulmonary Disease (COPD) is a heterogeneous disease with a variety of symptoms including, persistent coughing and mucus production, shortness of breath, wheezing, and chest tightness. As the disease advances, exacerbations, i.e. acute worsening of respiratory symptoms, may increase in frequency, leading to potentially life-threatening complications. Exposure to air pollutants may trigger COPD exacerbations. Literature predictive models for COPD exacerbations, while promising, may be constrained by their reliance on fixed air quality sensor data that may not fully capture individuals' dynamic exposure to air pollution. To address this, we designed a machine learning (ML) framework that leverages data from personal air quality monitors, health records, lifestyle, and living condition information to build models that perform short-term prediction of COPD exacerbations. The framework employs (i) k-means clustering to uncover potentially distinct patient sub-types, (ii) supervised ML techniques (Logistic Regression, Random Forest, and eXtreme Gradient Boosting) to train and test predictive models for each patient sub-type and (iii) an explainable artificial intelligence technique (SHAP) to interpret the final models. The framework was tested on data collected in 101 COPD patients monitored for up to 6 months with occurrence of exacerbation in 10.7% of total samples. Two different patient sub-types have been identified, characterised by different disease severity. The best performing models were Random Forest in cluster 1, with area under the receiver operating characteristic curve (AUC) of 0.90, and area under the precision/recall curve (AUPRC) of 0.7; and Random Forest model in cluster 2, with AUC of 0.82 and AUPRC of 0.56. The model interpretability analysis identified previous symptoms and cumulative pollutant exposure as key predictors of exacerbations. The results of our study set a premise for a predictive framework in COPD exacerbations, particularly investigating the potential influence of environmental features. The SHAP analysis revealed that the contribution of environmental features is not uniform across all subjects. For instance, cumulative exposure to pollutants demonstrated greater predictive power in cluster 1. The SHAP analysis also shown that overall clinical factors and individual symptomatology play the most significant role in this setup to determine exacerbation risk.

Chronic Obstructive Pulmonary Disease (COPD) represents a significant challenge to global health systems, contributing extensively to healthcare expenses, mortality, and morbidity worldwide. The Global Burden of Disease Study predicts COPD to become the fourth leading cause of death by 2030[1]. COPD is recognized as a heterogeneous disease comprising a variety of clinical conditions ranging from emphysema to chronic bronchitis, and is confirmed by irreversible airflow limitation[2,3]. People with COPD can have a variety of symptoms, including persistent coughing and mucus production, shortness of breath, wheezing, and chest tightness[4].

As the disease progresses, exacerbations, defined as acute worsening of respiratory conditions, may become more frequent, and life-threatening complications may develop[5,6]. Over time, this condition tends to deteriorate. One of the main characteristics of its decline is the reduction of the maximal flow rate measured during a forceful expiration following full inspiration. This measurement is also known as Peak Expiratory Flow rate (PEF). The

[1]Department of Information Engineering, University of Padova, Padova, Italy. [2]Environmental Research Group, MRC Centre for Environment and Health, Imperial College London, London, United Kingdom. [3]School of Public Health, Imperial College London, London, United Kingdom. ✉email: martina.vettoretti@unipd.it

exposure to air pollutants, especially Particulate Matter (PM10, PM2.5), nitrogen dioxide (NO2), ozone (O3), and carbon monoxide (CO), drives an inflammatory reaction in the lungs that can significantly increase the risk of lung function and/or symptomatic decline[3,7–10].

COPD exacerbations are predominantly caused by infection - viral or bacterial - but may also be triggered by air pollution. The heterogeneity of COPD patients' characteristics can be studied by descriptive statistical methods and unsupervised machine learning (ML) techniques, such as clustering that allows to identify coherent patient sub-types. In the literature, building models for patient sub-types as opposed to an overall model (i.e. global model) is widely used in the healthcare domain[11–15]. Specifically, in the COPD domain, clustering was mainly performed on clinical features[16], often overlooking behavioral factors except smoking habits[17,18]. Yet, the analysis of other features, such as living environment information and environmental exposures, which may have an important impact on COPD exacerbations, remains underexplored.

Several ML models have been proposed to predict the risk of exacerbation in COPD patients. Most of the models were designed to predict the number of exacerbations COPD subjects are going to experience in the next year[19,20]. Some other models predict the occurrence of exacerbations in a wide temporal window of upcoming days[21–23]. Various factors have been used as predictors, including patient-reported wellbeing, forced expiratory volume in one second (FEV1), C-reactive protein (CRP) levels[21], portable spirometry measurements, lung electronic auscultation data[24], lifestyle data, patient symptoms[23], influenza virus data[25], and exposure to air pollutants measured mostly from fixed air quality monitors[25,26].

Although fixed air quality monitors generally provide accurate measurements, they collect measurements at low spatio-temporal resolution. Using fixed air quality monitors, the individual exposure to air pollution is usually estimated considering the measurements of the monitor closest to the subject's residence. However, this approach does not allow to track the variability in the exposure due to the subject's activities and movements (e.g., when being at home, at the working place, when commuting, etc.). This limitation could be overcome by the use of personal air quality monitors (PAMs) that can be worn by the patients while moving through their activity spaces, thus allowing to monitor the personal exposure to air pollutants at a finer level, both outdoors and indoors.

Another limitation of the state-of-the-art predictive models of COPD exacerbations is that they usually do not rely on techniques to deal with the COPD heterogeneity, potentially missing the identification of sub-types among COPD patients that enables the development of customized predictive models.

In response to these challenges, our study introduces a ML framework that leverages data from PAMs, alongside lifestyle and health data, to develop new predictive models for the short-term prediction of COPD exacerbations. In particular, the framework performs the following analysis steps: (i) imputation of missing data by state-space Auto-Regressive Integrated Moving Average (ARIMA) models for pollutant time series and the Multiple Imputation by Chained Equations (MICE) algorithm for other data; (ii) feature engineering to extract relevant features from time series data collected over the past three days; (iii) k-means clustering to stratify COPD patients in different sub-types with similar characteristics; (iv) for each identified sub-type, training of three supervised ML models, i.e. Logistic Regression (LR), Random Forest (RF) and eXtreme Gradient Boosting (XGB), for the prediction of COPD exacerbation occurrence in the next day (the training pipeline includes feature selection and hyperparameter tuning); (v) evaluation of the model performances on out of bag testing samples; (vi) interpretation of the best performing models using an explainable artificial intelligence technique based on the SHapley Additive exPlanations (SHAP) values.

The framework is tested on data collected by a study that investigated the impact of individual air pollution exposure on lung function and symptoms in COPD patients living in London[27]. The study utilized PAMs worn by participants to gather direct pollutant exposure data, including pollutant concentration levels, temperature, humidity, and gas concentrations. Lifestyle factors and COPD-related symptoms were collected by enrollment questionnaires and daily logs throughout the study. The final dataset included 101 subjects monitored for up to six months. More details about the data and methods are reported in the "Methods" section.

Results

Patient stratification by clustering

The number of clusters was varied from k = 1 to k = 10. The choice of the optimal number of clusters k was performed using three metrics: the average silhouette width[28], the Dunn index[29] and the separation index[30]. The optimal number of clusters is often determined by maximizing these metrics, which reflect how well-separated the clusters are. As shown in Fig. 1, the average silhouette width tends to increase as k increases, with higher values for k = 9 or k = 10 compared to k = 2. This is likely due to the structure of the dataset, where we have multiple observations per subject. However, the silhouette width does not increase substantially beyond k = 2. Similarly, when increasing the number of clusters beyond two, minimal gains were obtained for the Dunn and the separation indexes, which for k = 2 were equal to 0.25 and 1.34 respectively. This suggests that while more clusters may slightly improve the separation of individual subjects, it does not contribute meaningfully to capturing main patterns across the dataset.

Moreover, an estimate of the cluster stability was evaluated computing the Jaccard's similarity over 100 bootstrap iterations[31] . The mean Jaccard similarity coefficient was 0.81 for Cluster 1 and 0.75 for Cluster 2, indicating high stability in the cluster assignments, as values above 0.75 are regarded to demonstrate strong cluster coherence.

Therefore, we concluded that two clusters represent a reasonable choice for the current problem.

Table 1 presents the main characteristics of the two identified groups, while Fig. 2 shows the distribution of the two clusters over the first two principal components obtained by the principal component analysis algorithm. As visible in Fig. 2, the two clusters appear well separated over the first two principal components that overall explain about 73.9% of the variance of the original data. Analysing the cluster characteristics reported in

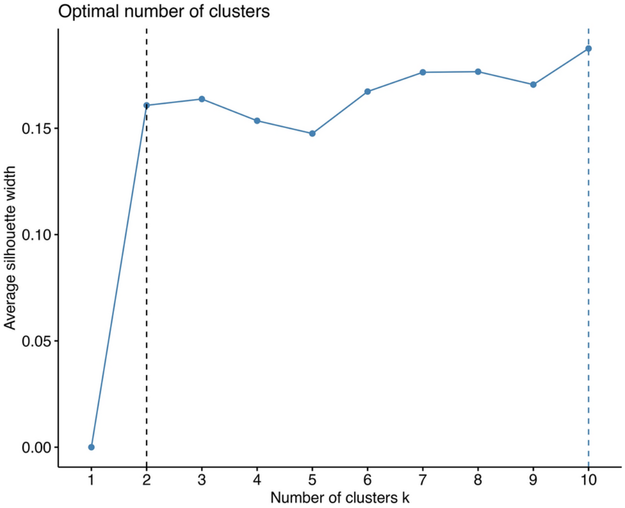

Fig. 1. Silhouette plot for k-means clustering with number of clusters k ranging from k = 1 up to k = 10.

Variable (Feature name)	Cluster 1	Cluster 2
	Subjects: 60	Subjects: 41
Age (Age)	71.00 (68.0, 76.00)	70.00 (65.0, 77.00)
Biological sex (Sex)		
Female (1)	33 (55.0%)	16 (39.0%)
GOLD Severity (COPDSeverity)		
Mild (1)	16 (26.7%)	2 (4.9%)
Moderate (2)	40 (66.7%)	3 (7.3%)
Severe (3)	1 (1.7%)	30 (73.2%)
Long-acting bronchodilator inhalers (LABA)	24 (40.0%)	32 (78.0%)
Long-acting muscarinic antagonists (LAMA)	37 (61.7%)	31 (75.6%)
Inhaled corticosteroids (ICS)	37 (61.7%)	34 (82.9%)
Former Smoker (FormerSmoker)	42 (70.0%)	35 (85.4%)
Quit Smoking [months] (QuitDurationMonths)	126.00 (6.0, 390.00)	96.00 (12.0, 168.00)
Car Owner (Car)	35 (58.3%)	29 (70.7%)
Days out/week (DaysOutPerWeek)	6.00 (4.9, 7.00)	5.50 (4.0, 7.00)
House Type (House_structure_type)		
Multi-level (1)	34 (56.71%)	13 (31.7%)
Compact Living (2)	20 (33.3%)	18 (43.9%)
Cooking Appliances (Cooking)		
Gas (1)	16 (26.7%)	24 (58.5%)
Electric (2)	32 (53.3%)	13 (31.7%)
	Total samples: 6260	Total samples: 4254
72-h Cumulative Exposure (Cum_exposure)		
CO [ppm]	798.82 (660.8, 993.59)	797.35 (664.5, 1,002.31)
NO2 [ppb]	26,262.93 (18,031.8, 38,102.75)	27,142.17 (17,654.3, 40,971.27)
O3 [ppb]	14,665.34 (9,985.3, 21,259.32)	16,413.34 (10,113.9, 25,202.68)
PM2.5 [µg/m3]	36,171.34 (23,828.2, 59,736.72)	40,923.11 (25,663.1, 66,022.33)
Sleep disturbances (Sleep_disturb)	710 (11.3%)	417 (9.8%)
Breathlessness (Breathlessness)	918 (15.0%)	1,119 (26.9%)
Other Symptoms (Other_symptoms)	1,454 (23.7%)	1,208 (29%)
Delta Peakflow [%] (DeltaPEF)	−5.56 (−12.5, 0.0)	−6.25 (−11.8, 0.0)
Exacerbation (Exacerbation)	912 (14.6%)	802 (18.9%)
Exacerbation prev. 30dd (PrevExacerbation_30gg)	3.47 (2.6, 4.30)	3.40 (2.5, 4.17)

Table 1. Descriptive statistics of the two clusters based on the k-means clustering analysis. The values represent median (interquartile range) or number of observations (%) for each variable.

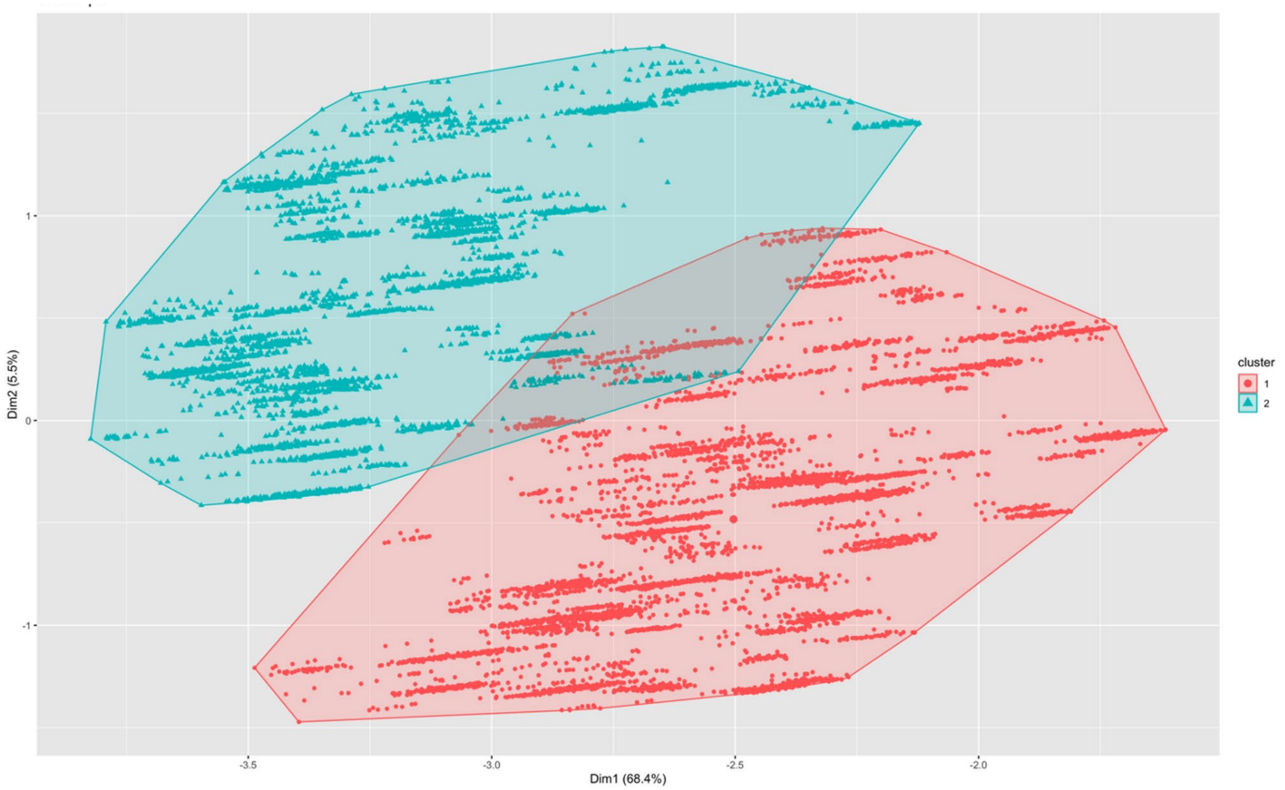

Fig. 2. Scatter plot showing the two clusters derived by the k-means algorithm, plotted in the first two principal components. Within this visualization, red dots mark data points belonging to the first cluster, while green triangles denote data points attributed to the second cluster.

Table 1, it can be observed that the first cluster, cluster 1, grouped subjects that had generally lower cumulative exposure levels, lower number of days with exacerbations and/or symptoms but a higher number of days with sleep disturbances. Moreover, cluster 1 shows a better general condition of COPD, with higher median values of PEF measurements and mild to moderate COPD severity. Instead, the second cluster, cluster 2, represents a group of patients with a more severe COPD level and that declared to go out less days in the week. Furthermore, the second cluster has a bigger proportion of subjects with gas cooking appliances.

Development of predictive models of COPD exacerbations

Table 2 presents the performances of the three ML models in terms of mean (standard deviation) across 10 test sets generated by bootstrap resampling. For more information about the predictive model training and the chosen metrics, please refer to the section "Predictive models training and assessment". The results are presented for each of the two clusters found with the k-means algorithm. Notably, given the high imbalance in the outcome (COPD exacerbations occurring in only 15% of the cases), the Area Under the Precision-Recall Curve (AUPRC) threshold of 0.15 is used as reference to account for the skewed distribution of the positive cases.

To further assess the impact of clustering on model performance, we conducted additional sensitivity analyses. We evaluated the models' performance without using the clustering labels to determine whether their inclusion significantly influenced the predictions. Therefore, the models were trained on the full dataset, and the results were compared to those obtained when clustering was included.

In cluster 1, the RF model achieved the highest AUC score of 0.90 (±0.05), with an AUPRC of 0.70 (±0.05), indicating excellent discriminatory power in distinguishing between positive and negative cases, where a positive case is an instance with a COPD exacerbation and a negative case is an instance without COPD exacerbation. RF also demonstrated high performance in terms of sensitivity (0.83 ± 0.05) and specificity (0.86 ± 0.05), suggesting its effectiveness in correctly identifying both true positives and true negatives. Similarly, the XGB model exhibited competitive performance, with an AUC of 0.89 (±0.04) and AUPRC of 0.65 (±0.10), alongside with sensitivity (0.87 ± 0.07) and specificity (0.82 ± 0.07). Both RF and XGB slightly outperformed the LR model, which had an AUC of 0.88 (±0.05) and an AUPRC of 0.65 (±0.10).

In contrast, the performance of the models in cluster 2 was generally lower compared to cluster 1. The RF model achieved the best performance with an AUC of 0.82 (±0.04) and AUPRC of 0.56 (±0.10), followed by XGB with an AUC of 0.81 (±0.04) and AUPRC of 0.5 (±0.1). LR performed slightly lower with an AUC of 0.79 (± 0.06) and AUPRC of 0.53 (± 0.10). Additionally, all models showed lower specificity in this cluster (values around 0.75), indicating that, even the non-linear models struggled with the high false positive rate (1-specificity) in this sub-type.

Model	AUC	AUPRC	Sensitivity	Specificity	G-mean
Cluster 1					
LR	0.88 (0.05)	0.65 (0.1)	0.76 (0.08)	0.87 (0.08)	0.81 (0.06)
RF	0.90 (0.05)	0.7 (0.05)	0.83 (0.05)	0.86 (0.05)	0.84 (0.01)
XGB	0.89 (0.04)	0.65 (0.1)	0.87 (0.07)	0.82 (0.07)	0.84 (0.04)
Cluster 2					
LR	0.79 (0.06)	0.53 (0.1)	0.75 (0.1)	0.78 (0.1)	0.76 (0.05)
RF	0.82 (0.03)	0.56 (0.07)	0.78 (0.1)	0.77 (0.1)	0.77 (0.04)
XGB	0.81 (0.04)	0.5 (0.1)	0.82 (0.06)	0.74 (0.07)	0.77 (0.03)
Overall (no clustering)					
LR	0.87 (0.03)	0.6 (0.09)	0.79 (0.06)	0.82 (0.06)	0.81 (0.03)
RF	0.87 (0.02)	0.62 (0.08)	0.85 (0.04)	0.79 (0.04)	0.82 (0.03)
XGB	0.87 (0.03)	0.59 (0.09)	0.83 (0.06)	0.81 (0.06)	0.82 (0.03)
Overall model evaluated on cluster 1 test sets					
LR	0.89 (0.04)	0.67 (0.1)	0.8 (0.07)	0.85 (0.07)	0.83 (0.04)
RF	0.90 (0.04)	0.7 (0.07)	0.87 (0.05)	0.83 (0.05)	0.85 (0.03)
XGB	0.89 (0.05)	0.66 (0.1)	0.83 (0.08)	0.86 (0.08)	0.84 (0.04)
Overall model evaluated on cluster 2 test sets					
LR	0.84 (0.04)	0.55 (0.09)	0.78 (0.07)	0.8 (0.07)	0.79 (0.03)
RF	0.83 (0.03)	0.55 (0.09)	0.81 (0.08)	0.76 (0.08)	0.78 (0.03)
XGB	0.82 (0.05)	0.52 (0.1)	0.8 (0.09)	0.76 (0.09)	0.78 (0.04)

Table 2. Out of bag (test set) performances comparison between the models for each cluster, without prior clustering (overall) and evaluating overall models on clustered test sets in terms of mean (standard deviation). Abbreviations: Logistic Regression (LR), Random Forest (RF), eXtremeGradientBoost (XGB).

When training the models on the whole training set, without cluster partition, and evaluating them on the whole test set, the performance scores presented intermediate values between those of Cluster 2 and Cluster 1. Also in this case the best performing model was RF, which achieved an AUC of 0.87 (±0.02), with balanced sensitivity (0.85 ± 0.04) and specificity (0.79 ± 0.04), while XGB and LR showed similar overall AUC scores of 0.87 (±0.03) but slightly lower AUPRC scores. The performance of the models trained on the whole training set was also assessed separately on the Cluster 1 test set and Cluster 2 test set (Table 2). In particular, the RF model achieves better performance in Cluster 1 compared to the overall population, especially considering the AUPRC score, which is 0.7 in Cluster 1 and 0.62 in the overall population, and the specificity, which is 0.86 in Cluster 1 and 0.79 in the overall population. On the other hand, the performance in Cluster 2 are worse compared to the overall population, suggesting that for this patient subgroup the prediction of exacerbations might be a more challenging task. These results suggest that clustering may highlight differences in predictive performance on patient sub-types with potentially distinct characteristics. Evaluating the overall model on cluster 1 and cluster 2 test sets revealed performance scores that closely aligned with the respective cluster-specific models, though slightly higher in most cases.

Additionally, we examined the effect of assigning cluster labels in the test set after model training. Starting with the centroids identified in the training set, test set samples were labeled by assigning them to the nearest cluster based on the minimum Euclidean distance from the centroids. The outcome event was then predicted using the model corresponding to the assigned cluster. The results were again consistent with the original analysis, demonstrating that the models performed similarly regardless of whether the clustering labels were assigned before or after training the models (Table 3).

Interpretation of best performing models on each cluster

Results reported in Table 2 evidence that clustering revealed a patient sub-type, i.e. Cluster 2, for which the prediction of exacerbations is more challenging. Although the training of cluster-specific models did not improve the model performance compared to training the models on the overall training set, we applied the SHAP technique on the best performing models in each cluster, to investigate possible differences in the predictive ability of the features in the two clusters. SHAP summary plots are shown in Fig. 3 for RF in cluster 1 and in Fig. 4 for RF in cluster 2. In these plots, each row corresponds to a feature, and each dot represents a specific instance from the training set. The color of the dot indicates whether the data instance has a low (purple) or high (ocher) value for the feature compared to its mean. Features are ranked by importance on the y-axis, with higher positions indicating greater importance. The position of a dot on the x-axis indicates its SHAP value, which measures the feature's impact on the model's prediction for that instance. The width of the plots in each row shows how many instances are associated with a specific SHAP value. Therefore, if instances with high feature values have large positive SHAP values, it suggests that the feature strongly influences the model outcome (i.e., higher feature values lead to higher probability of exacerbation). Further details on SHAP are reported in section "ML models explainability".

Model	AUC	AUPRC	Sensitivity	Specificity	G-mean
Cluster 1, with labels assigned based on training set centroids					
LR	0.88 (0.05)	0.64 (0.10)	0.76 (0.09)	0.87 (0.09)	0.81 (0.07)
RF	0.90 (0.05)	0.68 (0.10)	0.84 (0.08)	0.85 (0.08)	0.85 (0.05)
XGB	0.89 (0.04)	0.63 (0.10)	0.85 (0.07)	0.84 (0.07)	0.85 (0.04)
Cluster 2, with labels assigned based on training set centroids					
LR	0.79 (0.05)	0.55 (0.10)	0.76 (0.10)	0.75 (0.10)	0.75 (0.04)
RF	0.81 (0.03)	0.55 (0.07)	0.78 (0.10)	0.75 (0.10)	0.76 (0.03)
XGB	0.80 (0.04)	0.51 (0.10)	0.82 (0.08)	0.72 (0.08)	0.76 (0.03)
Cluster 1 + Cluster 2, with labels assigned based on training set centroids					
LR	0.84 (0.05)	0.59 (0.10)	0.76 (0.07)	0.81 (0.07)	0.78 (0.05)
RF	0.86 (0.03)	0.60 (0.10)	0.82 (0.05)	0.79 (0.05)	0.81 (0.03)
XGB	0.85 (0.04)	0.54 (0.10)	0.81 (0.06)	0.81 (0.06)	0.81 (0.03)

Table 3. Out of bag (test set) performances for models with cluster labels assigned after model training based on training set centroids. Performances are reported in terms of mean (standard deviation). Abbreviations: Logistic Regression (LR), Random Forest (RF), eXtreme Gradient Boosting (XGB).

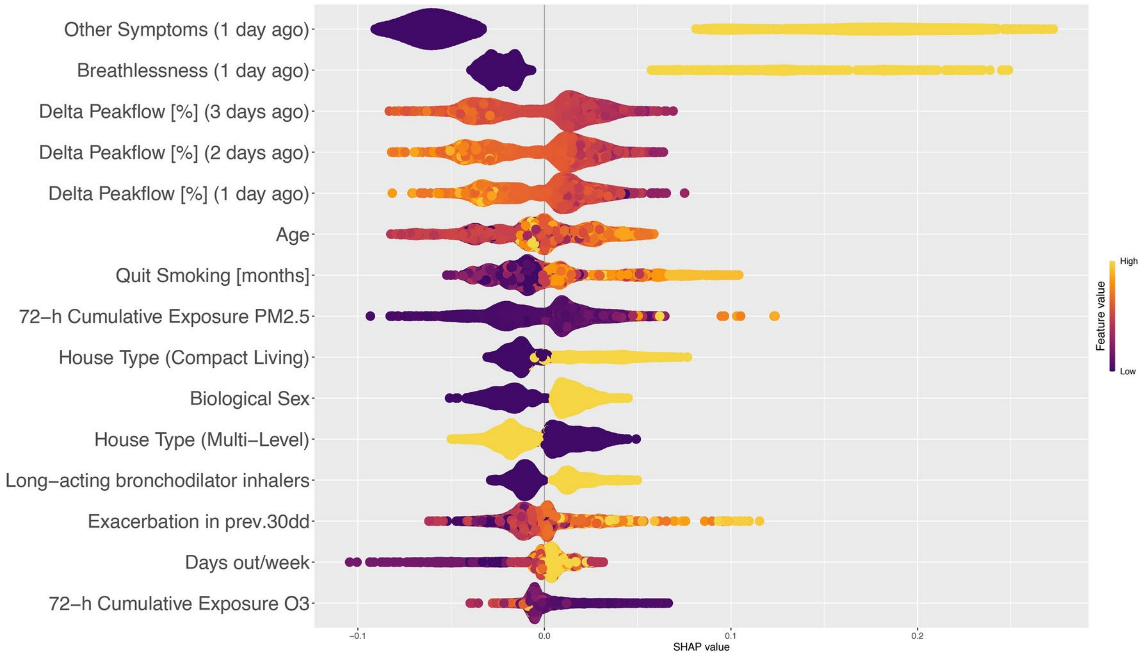

Fig. 3. SHAP values distributions for the 15 most impactful features for RF in cluster 1. Positive SHAP values (right side of the figures) are associated with an increase in exacerbation probability. Negative SHAP values are associated with a decrease in exacerbation probability. Distributions are color-coded, ocher portions are associated with high variable values while purple portions with low variable values.

Figure 3, displays the plot of the best RF model in cluster 1. Among the most impactful features, previous day symptoms turn out to be important predictors of exacerbations. Coherently, a higher frequency of exacerbations (PrevExacerbation_30gg) suggests a greater likelihood of experiencing subsequent exacerbations. Improvement of PEF in preceding days is associated with a lower exacerbation risk. Furthermore, the cumulative exposure to air pollutants in the previous three days helps predicting exacerbations: higher cumulative exposure to PM2.5 positively increases exacerbation probability. Other factors contributing positively to exacerbation risk include older age, biological sex, residence in compact spaces, the declared number of days spent outdoors per week and use of long-acting bronchodilator inhalers (LABA) medications. Conversely, a longer duration since quitting smoking is associated with a higher exacerbation risk. This could be due to an interaction with other factors (e.g., people that have stopped smoking since a longer time are older subjects and/or subjects with a more severe COPD condition).

The SHAP summary plot of cluster 2 (Fig. 4) confirms the findings obtained for cluster 1, although with some discrepancies. Notably, the features representing the cumulative exposure to pollutants in the previous three days indicate NO2 as the predominant pollutant source influencing exacerbation episodes in this cluster. Moreover,

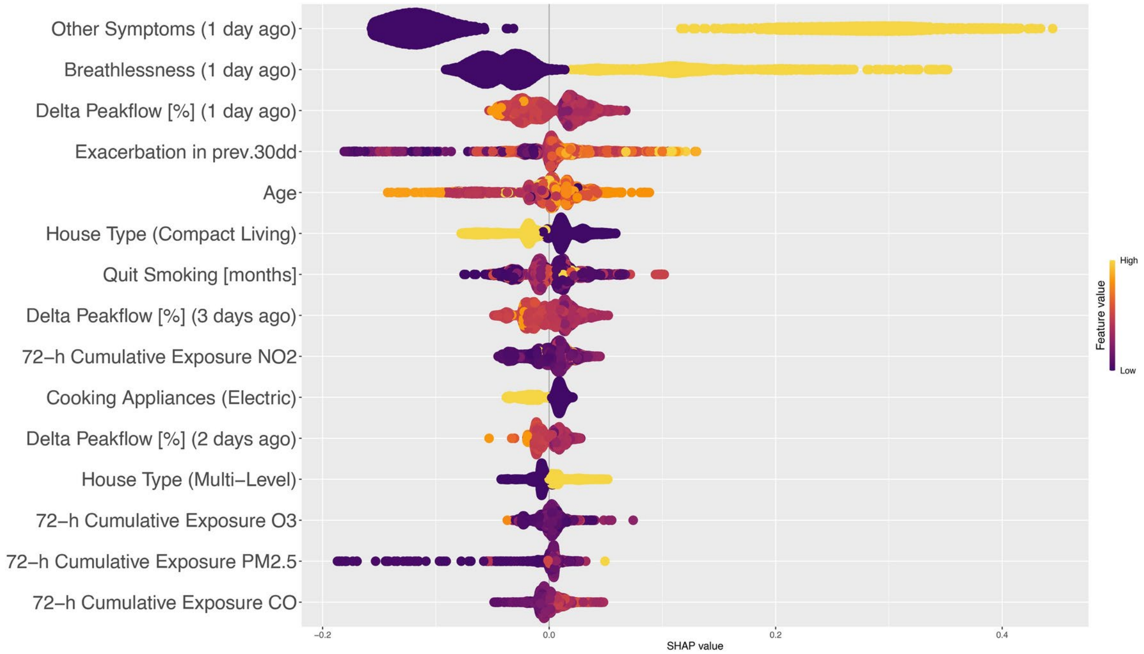

Fig. 4. SHAP values distributions for the 15 most impactful features for RF in cluster 2. Positive SHAP values (right side of the figures) are associated with an increase in exacerbation probability. Negative SHAP values are associated with a decrease in exacerbation probability. Distributions are color-coded, ocher portions are associated with high variable values while purple portions with low variable values.

this sub-type appears more sensitive to previous occurrences of exacerbations, which exhibit a notable impact on the risk. On the other hand, living in compact spaces with electric-based appliances are associated with a lower risk of exacerbations.

Moreover, we further quantified the contribution of the most impactful features by calculating the relative contribution (in percentage) of the mean absolute SHAP values with respect to the total contribution of all features. As expected, in cluster 1 symptoms occurring in the previous day (i.e., Other Symptoms (1 day ago) and Breathlessness (1 day ago)) together contribute to over 30% of the total prediction strength, while cumulative exposure to pollutants contributed for a total of 7%, with exposure to PM2.5 being the environmental feature with the highest contribution (5%) (Supplementary, Table S1). The results in cluster 2 indicate that previous day symptoms are the most influential factors (almost 50% of contribution), highlighting the importance of patient-reported outcomes in managing COPD in this subgroup. Environmental factors, apart from exposure to NO2 (contribution of 3%), did not contribute significantly to the prediction (contribution of 1% each), suggesting that this group might be slightly sensitive to NO2, but less sensitive to other pollutants compared to cluster 1 (Supplementary, Table S2).

Discussion

In our study, we proposed a ML framework which integrates data from PAMs with health data, lifestyle, and living environment information to develop models for making short-term predictions of exacerbation episodes in COPD patients. We employed a k-means clustering approach to uncover potentially distinct patient sub-types, thus enabling tailored predictive models to be developed. The k-means analysis identified two largely distinct sub-types: the first sub-type exhibited generally lower COPD severity and exposure profiles (cluster 1), while the second sub-type represented a cohort with poorer health outcomes who spend more time indoors (cluster 2).

For each identified sub-type, three supervised ML models (LR, RF, and XGB) were trained to predict the exacerbation occurrence in a day, using features extracted in the previous three days (Fig. 5). As evidenced in Table 2, our findings revealed variations in predictive accuracy among the linear and non-linear models, as well as across the two sub-types. Specifically, RF and XGB models performed generally better than LR in both cluster. In particular, the best performing model resulted RF. Regarding the comparison between clusters, while the prediction performance in cluster 1 was satisfactory, with good AUC (0.9) and AUPRC (0.7), as well as sensitivity and specificity values around 0.85, the performance achieved in cluster 2 was significantly lower, with a specificity value around 0.75. These results suggest that for some COPD patients, characterised by higher COPD severity, the prediction of exacerbations is more difficult, as predictive models are more likely to provide false positives.When training and evaluating models on the whole dataset, the performance scores remained relatively high, balancing the variability observed between the two clusters. Models trained on the whole training set and evaluated separately on each cluster test set showed similar performance to cluster-specific models. This suggest that in this population training cluster-specific models does not allow to improve the prediction performance compared to using a single model trained on the overall population. Nevertheless, the clustering analysis revealed the presence of two different patient sub-types, for which model performances are significantly different, being Cluster 2 the patient sub-type for which the prediction of exacerbations is more challenging.

SHAP summary plots were used to analyse the impact of each feature on model predictions in each cluster. Previous day symptoms, delta PEF values, and PM2.5 exposure were identified as significant indicators of exacerbation risk in the mild/moderate sub-types (cluster 1). In the most severe sub-type (cluster 2), NO2 was the main pollutant influencing exacerbation episodes, and previous exacerbations also had a notable impact on the exacerbation risk. Age, living environment and cooking appliances also played a role in predicting the exacerbation risk.

Regarding the limitations of our study, a first limitation concerns the definition of exacerbation events, which in the original observational study were self-reported by the patients using health logs. However, to mitigate the possible bias due to patient's subjectivity in reporting exacerbations, self-reported exacerbations underwent a validation by medical professionals that contributed to the robustness of the outcome[27]. Another limitation concerns the fact that exacerbation episodes are rare events, resulting in a highly unbalanced dataset. To address this issue, our pipeline employed the SMOTENC oversampling algorithm to balance the observations of each training set. Additionally, we analyzed both AUPRC and G-mean metrics (Eq. 1) which are recommended for unbalanced settings[32]. Finally, the non-uniform adherence to wearing the monitoring devices among participants may impact data completeness and correctness, although it is worth noting that COPD patients typically exhibit greater compliance with medical equipment compared to other population groups.

To address the issue of missing values, our pipeline employed state-space ARIMA models to reconstruct gaps in pollutant time series and the MICE algorithm for imputing the other missing data. Note that MICE works under the assumption that missing values are Missing At Random (MAR).This means that any systematic differences between the missing and observed data can be explained by the observed features included in the imputation model. If this assumption does not hold and the data are Missing Not At Random (MNAR), i.e. the probability of missingness depends on unobserved data, the imputation process could introduce bias, potentially leading to inaccurate estimates and weakened model performance. In our dataset, features potentially at risk of not satisfying the MAR assumption were covariates, symptoms, and PEF measurements; as patients might be less inclined to record their values when feeling sicker generally, acutely unwell, or less engaged. However, it is important to note that the overall percentage of missingness was very low. For the subject covariates, the missingness was less than 5%. Moreover, there were no missing values in the original PEF and symptom data. As a result, MICE was applied exclusively to the covariates before merging them with the rest of the dataset to have a minimal impact on the model's predictive capacity. While it is difficult to quantify the exact proportion of data that may be MNAR, the low rate of missing data reduces the likelihood of significant bias in missing data imputation. However, it remains a limitation of our approach that the imputation relies on the MAR assumption, and any deviation from this assumption could impact the accuracy of our results.

Despite these challenges, the results from our study set a premise for the development of models for the short-term prediction of COPD exacerbations, particularly regarding the potential influence of environmental features. However, the contributions of these features, as indicated by the SHAP analysis, revealed that their impact is not uniform across all subjects. For instance, while cumulative exposure to pollutants demonstrate some predictive power, enhancing the performance metrics in cluster 1, the models also show that overall clinical factors and individual symptoms play the most significant role in this setup to determine exacerbation risk.

The results obtained by the present study are somehow in line with the results obtained by other literature studies that used fixed air quality sensors to measure the exposure to air pollution, although a direct comparison cannot be done because datasets with different features and collected in different settings were used. For example, Jo et al.[25] identified several significant predictors of acute exacerbations (i.e., outpatient visit with prescription of systemic steroids or visit to an emergency room and/or hospital admission), e.g., age, sex, smoking status, comorbidities, FEV1 value, disease specific medication usage, frequency of previous exacerbation, fixed air quality sensor data (NO2 and PM2.5), humidity and diurnal temperature data. Most of these factors resulted significant predictors of exacerbations also in the present study. Regarding prediction performance, Jo et al.[25] achieved similar results to the one obtained in this study for cluster 1 (AUC exceeding 0.9 in the internal validation set and 0.7 in the external validation set, using various ML models, e.g., RF, XGB, LightGB, and Mixed Effect Random Forest). In another recent study, Ratcliff et al.[26] explored the impact of air quality data, extracted from fixed monitoring stations, on RF models that predict COPD exacerbations requiring medical care. Their findings showed a slight improvement with air quality related predictors, reaching an AUC value of 0.85, and their net benefit curves suggested greater clinical utility when considering this data. Finally, Wu et al.[23] integrated home air quality sensors and wearable devices to monitor lifestyle and indoor environments and their impact on COPD exacerbations. Their best-performing deep neural network achieved good sensitivity (94%) and specificity (90.4%), with an AUC exceeding 0.9. However, limitations in environmental data collection (limited to user bedrooms) were acknowledged to potentially degrade prediction results.

Despite significant interest in the impact of air pollution on chronic respiratory conditions, few studies have attempted to build a predictive model of short-term exacerbations of COPD. The present study provides new evidence in this field, especially leveraging PAMs measurements and patient's sub-typing for customizing the model. The ML framework, as proposed in this work, once refined, offers multiple potential pathways for integration into clinical practice. The main goal of the framework is to predict exacerbations in the short term, providing clinicians with actionable information to intervene early. This could be integrated into clinical decision support systems (CDSS), alerting physicians or patients when exacerbation risk is high. This early warning could prompt timely interventions, such as medication adjustments, environmental exposure reduction, or increased monitoring, potentially preventing exacerbations from occurring or mitigating their severity. Moreover, by employing SHAP, clinicians may better understand and explain to patients the factors driving their risk of exacerbation. This may help shared decision-making, where patients are more informed and engaged in managing their condition. For example, clinicians can explain the impact of air pollution and personal behavior (i.e. cooking habits) on exacerbation risk, making it easier to align patient behavior with recommended interventions. Finally, from a broader clinical perspective, healthcare systems can use this model for population-level predictions. High-risk subgroups, such as those with severe disease or living in high-pollution areas, can be identified and monitored more closely. This allows for more efficient allocation of healthcare resources, targeting those patients most in need of intensive management or early intervention.

Future work includes the exploration of alternative features and feature engineering approaches (e.g., collecting more information about sleep and its influence on exacerbation, extracting different summary statistics, etc.), as well as modifying the framework itself (e.g. introducing L1/L2 regularized LR models and survival analysis models, implementing different approaches for managing class imbalance, such as class weighting, etc.), and finally the validation of the proposed ML framework on data collected in other populations with chronic respiratory diseases (e.g., asthma). A sensitivity analysis could also be performed to better understand the impact of using PAMs data in predictive models of COPD exacerbations. Additionally, given the low frequency of exacerbation events, developing robust predictive models remains a challenge that necessitates further investigation with larger datasets.

Methods

Dataset

The COPE study was designed to explore the relationship between individual air pollution exposure levels and lung function, symptoms of COPD, and the occurrence of exacerbations among a group of 101 COPD patients. This was the first COPD study utilizing direct measurements of personal exposure to various pollutants, collected by PAMs over a period of six months, instead of relying on indirect measurements or predictive models. This extended monitoring period allowed measurements to be collected in a wide range of conditions and activities.

Recruitment was conducted via general practitioners who accessed patients through the Clinical Practice Research Datalink (CPRD), a database of anonymized general practice records providing a wealth of ongoing primary healthcare data. Participants were then invited to a clinic where they were equipped with a PAM to use both at home and during their daily activities.

PAMs, custom-made for the study, are wearable devices designed to monitor various environmental factors. They can be worn around the waist or over the shoulder. PAMs measure PM ($\mu g/m^3$), temperature ($^\circ$C), humidity (%), and gases including CO, NO, NO2, and O3 (ppb), simultaneously. To quantify gases, electrochemical sensors based on amperometric methods are used, while PM masses are estimated using a miniaturized optical particle counter. This counter records particle counts in 16 sizes (bins) ranging from 0.35 to over 17 μm. Data collected by PAMs were automatically uploaded to a secure server via the charging base station, typically every 30 hours.

At the beginning of the study, a detailed questionnaire was used to gather data on lifestyle, such as days out per week, and living environment information including dwelling type, kitchen appliance types (such as wood-burning stoves, gas, or electric cookers), and car ownership. Throughout the study period, participants collected COPD-status-related daily logs, detailing symptoms, sleep quality, and PEF rates, which were determined by averaging three consecutive PEF measurements. Further details on the data collection methods and the specific instruments utilized in this research are available in[8,27,33].

The cohort comprised subjects with an average age of 70.8 ± 7.95 years, of whom 48% were females. According to the GOLD COPD severity criteria[34], 60.4% of the participants were classified with mild to moderate severity, whilst 40.6% as severe to very severe. A significant portion of subjects, 76.2%, were former smokers, with an average time since quitting smoking of 144.3 ± 150.4 months. Regarding living environment information, 46.5% resided in multi-level buildings (as detached, semi-detached or terrace houses), 37.6% in compact living units (flats, maisonettes), and the remainder in single-level homes (bungalows, mobile homes, cottages). Cooking appliances were primarily gas (39.6%) and electric (44.5%) based.

Throughout the study, exacerbations incidents were reported on 1,714 samples (10.7%), breathlessness on 2,037 samples (19.8%), other symptoms (wheeze, sputum and cough) on 2,662 samples (25.9%), and sleep disturbances on 1,127 samples (10.7%). The median for PEF was 240 L/min (interquartile range, IQR: 160, 320). The median 72-hour pollutant concentration levels were 0.16 ppm (IQR: 0.1, 0.19) for CO, 4.67 ppb (IQR: 3.4, 6.74) for NO2, 2.9 ppb (IQR: 1.9, 4.5) for O3, and 5.8 µg/m3 (IQR: 3.9, 9.1) for PM2.5.

Imputation

The reliability of the research outcomes depends on the quality and integrity of the underlying data. One of the major issue is represented by missing data. For PAMs data, missing entries often arise from sensor malfunctions, human errors in using the device, or power issues. On the other hand, missingness in personal/health data may be mainly caused by forgetfulness and the patient burden in data logging. To mitigate these gaps, we developed a robust imputation pipeline.

Missing data on each PAM time-series and for each subject were independently imputed as follows. Initially, days with missing data exceeding six hours were excluded from the analysis, as feature extraction for these days was considered not reliable. For the remaining days, small portions of missing data, spanning up to 10 minutes, were finely imputed using a state-space ARIMA model[35,36] and the Kalman filter. In particular, the state-space ARIMA model allows to model the structure and temporal dependencies within a time-series. Complementing this, the Kalman filter acts as an estimator, dynamically adjusting its predictions, hence missing data imputation is performed based on both the observed data and the inherent uncertainty in the model. Parameters of the ARIMA model were estimated for each participant and PAM time-series. Larger gaps up to six hours were imputed using the mean values from the gap adjacent measurements.

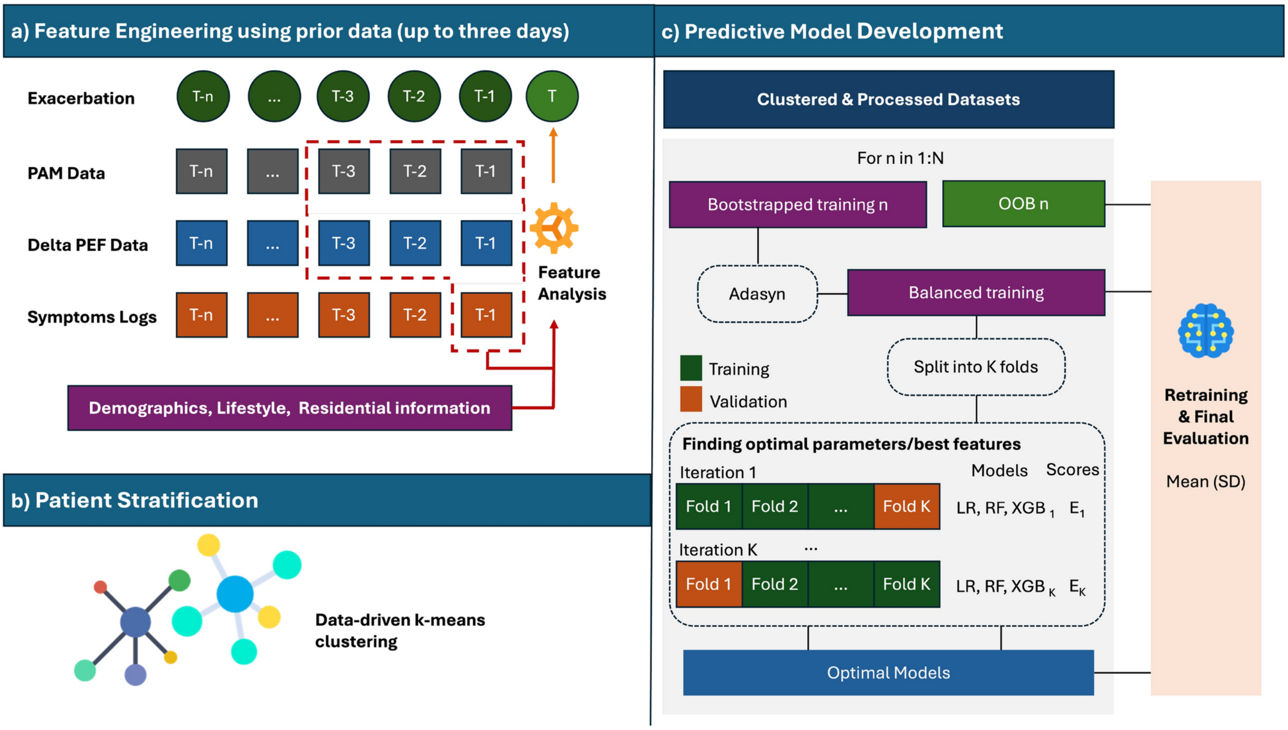

Fig. 5. ML framework for predicting COPD exacerbations. Lagged data on air pollution (PAM), lung function (PEF) and symptoms are combined with demographics, lifestyle, and living environment data and preprocessed (**a**). Then, processed data is clustered (**b**), and used to train ML models via a robust pipeline (**c**).

Missing values on other personal data, i.e., data on health, lifestyle, or living environment information, were imputed using the MICE method[37]. MICE imputes the missing values of each variable using the observed values of other features. This prediction is repeated several times, "chaining" together the imputations from each variable until convergence is reached. The overall missingness for these type of features was very low, typically below 5%. Moreover, crucial features, such as previous symptoms and PEF measurements, were not subject to imputation, as they remained fully observed.

Feature engineering

Each outcome instance in the dataset was defined on a daily basis within the six-month study period. Thus, for each patient, we generated multiple instances corresponding to each day, with features reflecting summary statistics for PAM over a three-day window, along with static covariates (such as age etc.), lagged delta PEFs and previous day symptoms.

PEF measurements were processed to compute the delta PEF feature, which is defined as the percentage difference between a measured PEF value and the individual's personal best PEF value. The personal best PEF value was determined as the highest PEF measurement among the first five PEF measurements collected during an healthy period, i.e., a period without exacerbation, changes in medication, or any symptoms recorded. The overall median value of the delta PEF feature was -6% (IQR: -12%, 0). Delta PEF features were lagged up to three days prior, such that for each outcome instance, we included three lagged delta PEF values: deltaPEF_lag.1 (from 1 day ago), deltaPEF_lag.2 (from 2 days ago), and deltaPEF_lag.3 (from from 3 days ago).

From the PAMs time-series, several summary statistics were calculated over a sliding three-day window prior to each outcome event, including median, standard deviation, minimum-maximum range, median absolute deviation, i.e., the median of the absolute deviations from the median, and cumulative exposure, which was defined as the integral of the signal in the 3-day (72-hour) time window computed using the trapezoidal rule. As an example, for each outcome instance (e.g., the outcome on Sept 30th), we calculated the median, standard deviation, minimum-maximum range, median absolute deviation, and cumulative exposure using data collected from the three days preceding the event (i.e., from Sept 27th to Sept 29th). The median of cumulative exposure resulted 798.1 ppm (IQR: 662.2, 996.79) for CO, 26,618.50 ppb (IQR: 17,883.6, 39,305.41) for NO2, 15,224.08 ppb (IQR:10,032.5, 2,711.51) for O3, and 38,026.78 µg (IQR: 24,454.3, 62,455.99) for PM2.5.

To prevent model bias, features with low variability, high correlation with other features, or a high portion of missing data were excluded from the following analyses. Precisely, features having more than 90% of the data with the same value were excluded. Consequently, categorical features related to visits, namely general practitioner visits and hospital visits, were removed. Then, a correlation analysis was conducted to identify features highly correlated with other features (Pearson's correlation coefficient larger than 0.8).

In total, we constructed a pre-processed dataset comprising 10,514 instances from 101 subjects, which were used for both k-means clustering and model training.

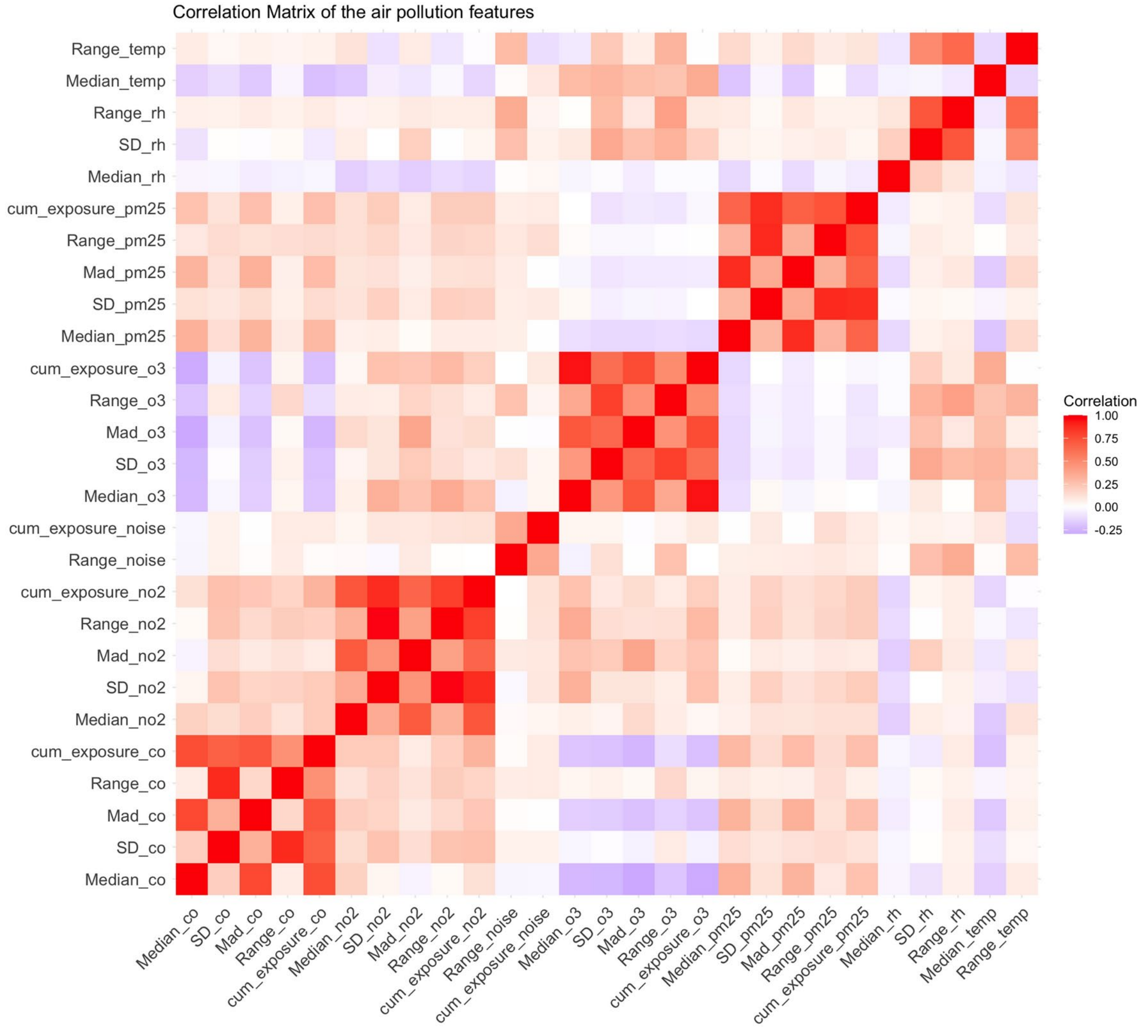

Fig. 6. Pearson's correlation matrix for the features extracted from PAMs time-series. The following suffixes are used to denote specific environmental parameters: _temp for temperature, _rh for relative humidity, _pm25 for particulate matter (2.5 μg), _o3 for ozone, _noise for ambient noise, _no2 for nitrogen dioxide, _co for carbon monoxide. In the visualization, the color red is associated with a positive correlation whereas the color purple denote a negative correlation.

As illustrated in Fig. 6, the correlation among PAMs features was generally low, except for features extracted from the same pollutant time-series which had a correlation coefficient >0.8. Therefore, we decided to consider a single feature for each pollutant. Taking into account the information content and interpretability of each feature, we decided to consider the cumulative exposure feature for the following analyses, as this feature provided a comprehensive representation of air pollution dynamics while mitigating redundancy. In fact, the cumulative exposure feature takes into account both the average of the signal and its fluctuations.

Clustering analysis

K-means clustering was performed for partitioning subjects into groups based on their similarities (Fig. 5b). This clustering algorithm starts by randomly selecting k data points from the dataset to serve as initial cluster centroids (centers of the clusters). Each data point in the dataset is assigned to the nearest centroid based on a distance metric (in our work, the Euclidean). After the assignment step, the centroids are recalculated as the mean of all data points assigned to each cluster, moving them to the center of their respective clusters. These steps are repeated iteratively until the centroids are no longer significantly modified.

K-means was applied on the preprocessed data considering the following features: cumulative exposure to PM2.5, CO, NO2 and O3; sleep disturbances, symptoms (breathlessness and other), delta PEFs (up to three days before), frequency of exacerbations, COPD severity, smoking history (former smoker and months since quitting), living conditions (house type, cooking appliances, car ownership, days out per week), demographics (age and biological sex), and medication type (LABA, LAMA, ICS). Minimum-maximum normalization of numerical features and one-hot encoding of multi-level categorical features were performed to ensure that the features had a uniform scale. To identify the optimal number of clusters, parameter k was varied from 1 to 10.

For each k value, three performance metrics were computed: Silhouette score, which measures clustering quality by comparing a sample's average distance within its cluster (a) to the nearest other cluster (b), and calculated as (b - a) / max(a, b)[28]; the Dunn index, which measures the ratio of the smallest distance between observations not in the same cluster over the largest distance in between data points of the same cluster[29]; and the separation index, which is the magnitude of the gap or sparse area between a pair of clusters (or cluster distributions) along the specified projection direction[30]. In general, higher values of these metrics indicate more separated clusters.

To evaluate the stability of our clustering solution, we valuated the Jaccard's similarity using the 'clusterboot' function from the R's fpc package[38], performing 100 bootstrap iterations. The Jaccard coefficient[39], calculated as the ratio of shared points between two clusters compared to the total points, shows how similar the clusters are. To assess stability, clusters are repeatedly compared across bootstrapped clustered samples, and the average Jaccard coefficient is used as a stability score. Higher values indicate more stable clusters. This method is non-parametric, making it suitable for general clustering methods, even those that do not rely on Euclidean distances.

After performing the clustering analysis, we partitioned the unnormalized dataset accordingly, creating a training and a test set for each cluster.

Predictive models training and assessment

For each identified patient sub-type, as well as in the overall population without clustering, we developed and tested ML models for the short-term prediction of exacerbations. Additionally, we tested an alternative scenario where the cluster labels for the test set were assumed to be unknown. Cluster centroids were computed from the training sets, and each test instance was assigned to the closest cluster based on the Euclidean distance to these centroids, after which the corresponding cluster-specific model was applied.

Initially, N=10 iterations of data resampling with replacement, namely bootstrapping, were done to create different training sets and test sets. Data resampling was performed by patient, so that data of the same patient could not appear both on the training set and the test set. In particular, for each iteration, data of 101 patients randomly sampled with replacement were included in the training set, the rest of patients were included in the test set. Min-max normalization was then applied specifically to each training set, and the same scaling parameters were used to normalize the test sets, thus ensuring consistency and reliability in the model evaluation. For each partition, a binary outcome was defined for each day of monitoring, as the occurrence of an exacerbation (positive class) or not (negative class). Outcome prediction was performed considering the features used by the k-means clustering, with cumulative exposure computed on the previous 3 days, delta PEF features lagging previous three days data, frequency of exacerbations computed from the previous 30 days up to the current day, and additional health and lifestyle covariates. The outcome had unbalanced classes, with only roughly 15% of positive values in each partition.

To address the problem of class imbalance, we applied the Synthetic Minority Over-sampling Technique for Nominal and Continuous features (SMOTENC)[40]on the training samples. In datasets where one class is substantially less represented than another, known as class imbalance, ML models may struggle to accurately learn patterns from the minority class. SMOTENC tackles the problem by generating synthetic "hard to learn" examples specifically for the minority class. Designed as an adaptation of the SMOTE algorithm specifically for datasets containing both categorical and continuous (numerical) features, it works by generating synthetic samples for the minority class to balance the class distribution in imbalanced datasets. It does this by interpolating new synthetic samples between existing minority class samples, but with special handling for categorical features to avoid unrealistic values. For a sample in the minority class, SMOTENC finds the k-nearest neighbors within the minority class. The nearest neighbors are determined based on both continuous and categorical features. For continuous features, traditional distance metrics like Euclidean distance are used. For categorical features, a Hamming distance (i.e., 0 if the categories are the same, 1 if they are different) is typically applied. The synthetic sample is then generated as follows: for continuous features, it creates new noisy values along the "line" connecting the two points; for categorical features, SMOTENC selects the categorical value from either the original sample or one of its nearest neighbors based on a majority vote or random selection.

After the balancing step, three ML models were trained on the balanced training set: a linear model, LR, and two non-linear models, RF and XGB. Each model was trained using an optimization pipeline comprising a feature selection and/or hyperparameter tuning phase to avoid overfitting and optimize model performance.

Precisely, for the LR, the Recursive Feature Elimination (RFE) was employed with 100 iterations. In each iteration, training data was divided in two subsets, referred to as subset training and subset validation. For each iteration, the RFE works by iteratively removing the least important features from the subset training until the desired number of features, or optimal performance, is reached. To do so, the algorithm starts by training a model on the entire set of features. After training the model, it ranks the features based on their contribution to the model's performance, here measured by the area under the receiver operating characteristic curve (AUC) on the subset validation. The least important feature (or features) is then removed from the dataset. This is done iteratively until all the features are removed.

For the RF and XGB a grid search approach within a 10-fold cross-validation was performed for hyperparameter tuning. For RF, we optimized mtry, i.e., the number of features to possibly split at in each node, ntree, denoting the number of trees in the forest, and maxdepth, i.e., the maximum depth of individual trees. The grid search strategy for RF hyperparameter choice was based on the following values: mtry from 1 to 15, ntree at values of 100, 250, and 500, while maxdepth varied between 3, 5, and 7. For XGB, the hyperparameters being tuned were nrounds, i.e., the maximum number of iterations, eta, which is the step size of each boosting step, max depth, defined as the maximum depth of the tree, colsample by tree, defined as the subsample ratio of columns when constructing each tree, and subsample, i.e., the subsample ratio of the training instance. To better clarify the latter, setting a subsample to 0.5 means that XGB randomly collected half of the data instances to grow trees. The assigned values on the grid were: 10, 50 or 100 for nrounds; 0.001, 0.005, 0.01 for eta; 2, 3, 5 for max depth; 0.5, 1 for subsample; and 0.5, 0.7 and 1 for colsample by tree. The performance metric considered for selecting the optimal models for each training/test set partition was the AUC computed on the validation folds. Figure 5c outlines the procedure.

Additionally, the prediction performance of the optimal models was evaluated on the test sets by the following metrics: sensitivity (recall), specificity (true negative rate), AUC, Area Under the Precision Recall Curve (AUPRC), and Kubat's geometric mean (Gmean)[32], which is defined as in the following equation:

$$Gmean = \sqrt{\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}} \tag{1}$$

where TP represents the number of true positives, FN represents the number of false negatives, TN represents the number of true negatives, and FP represents the number of false positives.

ML models explainability

Because of their complexity, non-linear models are more difficult to interpret. Nevertheless, several techniques are available for interpreting ML models. For the interpretability of our models and to analyze how different features influence the predicted outcome, we investigated the SHAP values. SHAP decomposes the output of a model by the sums of the impact of each feature. Precisely, calculates a value that represents the contribution of each feature to the model outcome, hence representing the importance of each feature to explain the result of the models. To do so, for each model input feature, its SHAP values were computed by interchanging its value with those from other input features randomly selected from the training set, as detailed in Lundberg and Lee[41]. This process involved interpreting each training set input independently, against a background set composed of N elements (in this study set to 500) randomly chosen from the training set itself. Finally, the obtained SHAP values were visualised using the SHAP summary plot. The summary plot generated from the SHAP values illustrates how each feature influences the model's predictions across all instances. The y-axis lists the most impactful features, while the x-axis represents the SHAP values, indicating the strength and direction of each feature's contribution to the predictions. Each point on the plot corresponds to a single prediction, colored according to the feature value (ocher for high and purple for low). This coloring helps visualize how high or low values of a feature affect the increasing risk of the exacerbation event.

Data availability

The data that support the findings of this study are available from Imperial College London but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the originating study PI (Barratt).

Received: 24 July 2024; Accepted: 31 December 2024
Published online: 18 January 2025

References

1. Quaderi, S. & Hurst, J. The unmet global burden of COPD. *Glob. Health Epidemiol. Genom.***3**. https://doi.org/10.1017/gheg.2018.1 (2018).
2. Barnes, P. Chronic obstructive pulmonary disease. *N. Engl. J. Med.* **343**, 269–80. https://doi.org/10.1056/NEJM200007273430407 (2000).
3. Pauwels, R. et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* **163**, 1256–76. https://doi.org/10.1164/ajrccm.163.5.2101039 (2001).
4. American Thoracic Society. Standards for the diagnosis and care of patients with chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* **152**, S77-121 (1995).
5. Suissa, S., Dell'Aniello, S. & Ernst, P. Long-term natural history of chronic obstructive pulmonary disease: severe exacerbations and mortality. *Thorax* **67**, 957–963 (2012).
6. Sutherland, E. & Cherniack, R. Management of chronic obstructive pulmonary disease. *N. Engl. J. Med.* **350**, 2689–97. https://doi.org/10.1056/NEJMra030415 (2004).
7. Choi, J. et al. Harmful impact of air pollution on severe acute exacerbation of chronic obstructive pulmonary disease: particulate matter is hazardous. *Int. J. Chron. Obstruct. Pulmon. Dis.* **13**, 1053–1059. https://doi.org/10.2147/COPD.S156617 (2018).
8. Evangelopoulos, D. et al. Personal exposure to air pollution and respiratory health of COPD patients in London. *Eur. Respir. J.*[SPACE]https://doi.org/10.1183/13993003.03432-2020 (2021).
9. Foster, W., Brown, R., Macri, K. & Mitchell, C. Bronchial reactivity of healthy subjects: 18–20 h postexposure to ozone. *J. Appl. Physiol.* **1985**(89), 1804–10. https://doi.org/10.1152/jappl.2000.89.5.1804 (2000).
10. Boehm, A. et al. COPD exacerbations are related to poor air quality in Innsbruck: A retrospective pilot study. *Heart Lung* **50**, 499–503. https://doi.org/10.1016/j.hrtlng.2021.02.012 (2021).
11. Hoogendoorn, M., el Hassouni, A., Mok, K., Ghassemi, M. & Szolovits, P. Prediction using patient comparison vs. modeling: A case study for mortality prediction. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Orlando, 2464–2467. https://doi.org/10.1109/EMBC.2016.7591229 (2016).
12. Baytas, I. M. et al. Patient subtyping via time-aware LSTM networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*, 65–74. https://doi.org/10.1145/3097983.3097997 (Association for Computing Machinery, 2017).
13. Suresh, H., Gong, J. J. & Guttag, J. V. Learning tasks for multitask learning: Heterogenous patient populations in the ICU. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '18)*, 802–810, https://doi.org/10.1145/3219819.3219930 (Association for Computing Machinery, 2018).
14. Galagali, N. & Xu-Wilson, M. Patient subtyping with disease progression and irregular observation trajectories (2018). arXiv:1810.09043.

15. Che, C. et al. An RNN architecture with dynamic temporal matching for personalized predictions of parkinson's disease. In *Proceedings of the 2017 SIAM International Conference on Data Mining (SDM)*, 198–206. https://doi.org/10.1137/1.9781611974973.23 (SIAM, 2017).
16. Kim, S. et al. A cluster analysis of chronic obstructive pulmonary disease in dusty areas cohort identified three subgroups. *BMC Pulm. Med.* **17**, 209. https://doi.org/10.1186/s12890-017-0553-9 (2017).
17. Burgel, P. et al. Clinical COPD phenotypes: a novel approach using principal component and cluster analyses. *Eur. Respir. J.* **36**, 531–539 (2010).
18. Siedlinski, M. et al. Genome-wide association study of smoking behaviours in patients with COPD. *Thorax* **66**, 894–902 (2011).
19. Adibi, A. et al. The acute COPD exacerbation prediction tool (ACCEPT): a modelling study. *Lancet Respir. Med.* **8**, 1013–1021. https://doi.org/10.1016/S2213-2600(19)30397-2 (2020).
20. Zeng, S., Arjomandi, M., Tong, Y., Liao, Z. & Luo, G. Developing a machine learning model to predict severe chronic obstructive pulmonary disease exacerbations: Retrospective cohort study. *J. Med. Internet Res.* **24**, e28953. https://doi.org/10.2196/28953 (2022).
21. Patel, N., Kinmond, K., Jones, P., Birks, P. & Spiteri, M. Validation of copdpredict^{T}M: Unique combination of remote monitoring and exacerbation prediction to support preventative management of copd exacerbations. *Int. J. Chron. Obstruct. Pulmon. Dis.* **16**, 1887–1899. https://doi.org/10.2147/COPD.S309372 (2021).
22. Samp, J. et al. Predicting acute exacerbations in chronic obstructive pulmonary disease. *J. Manag. Care Spec. Pharm.* **24**, 265–279 (2018).
23. Wu, C. et al. Acute exacerbation of a chronic obstructive pulmonary disease prediction system using wearable device data, machine learning, and deep learning: Development and cohort study. *JMIR Mhealth Uhealth* **9**, e22591. https://doi.org/10.2196/22591 (2021).
24. Yin, H. et al. A machine learning model for predicting acute exacerbation of in-home chronic obstructive pulmonary disease patients. *Comput. Methods Programs Biomed.* **246**, 108005. https://doi.org/10.1016/j.cmpb.2023.108005 (2024).
25. Jo, Y., Han, S. & Lee, D. Development of a daily predictive model for the exacerbation of chronic obstructive pulmonary disease. *Sci. Rep.* **13**, 18669. https://doi.org/10.1038/s41598-023-45835-4 (2023).
26. Ratcliff, G. et al. Integrating clinical and air quality data to improve prediction of COPD exacerbations. *AMIA Annu. Symp. Proc.* **2023**, 1209–1217 (2024).
27. Moore, E. et al. Linking e-health records, patient-reported symptoms and environmental exposure data to characterise and model copd exacerbations: protocol for the cope study. *BMJ Open* **6**, e011330. https://doi.org/10.1136/bmjopen-2016-011330 (2016).
28. Roussew, P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
29. Dunn, J. C. Well-separated clusters and optimal fuzzy partitions. *J. Cybern.*[SPACE]https://doi.org/10.1080/03610918.2014.911894 (1974).
30. Qiu, W. L. & Joe, H. Separation index and partial membership for clustering. *Comput. Stat. Data Anal.*[SPACE]https://doi.org/10.1016/j.csda.2004.09.009 (2006).
31. Hennig, C. Cluster-wise assessment of cluster stability. *J. Comput. Stat. Data Anal.* **52**, 258–271 (2007).
32. Kubat, M. & Matwin, S. Addressing the curse of imbalanced training sets: One-sided selection. In *Proc 14th International Conference on Machine Learning*, vol. 97, 179–186 (1997).
33. Quint, J. et al. Recruitment of patients with chronic obstructive pulmonary disease (COPD) from the clinical practice research datalink (CPRD) for research. *NPJ Prim. Care Respir. Med.* **28**, 21. https://doi.org/10.1038/s41533-018-0089-3 (2018).
34. Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease (2020 report). https://goldcopd.org/ (2020).
35. de Jong, R., van Buuren, S. & Spiess, M. Multiple imputation of predictor variables using generalized additive models. *Commun. Stat. Simul. Comput.*[SPACE]https://doi.org/10.1080/03610918.2014.911894 (2014).
36. Afrifa-Yamoah, E., Mueller, U. A., Taylor, S. M. & Fisher, A. J. Missing data imputation of high-resolution temporal climate time series data. *Meteorol. Appl.*[SPACE]https://doi.org/10.1002/met.1873 (2020).
37. van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**, 1–67 (2011).
38. Hennig, C. fpc: Flexible procedures for clustering. R package version 2.2-11 (2023).
39. Jaccard, P. Distribution de la florine alpine dans la bassin de dranses et dans quelques regiones voisines. *Bull. Soc. Vaud. Sci. Nat* **37**, 241–272 (1901).
40. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
41. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 1–10 (2017).

Acknowledgements

M.A. acknowledges the support of MUR (Italian Ministry of University and Research) under the PON initiative ex DM 1061.

Author contributions

M.A. and M.V. developed the algorithm; M.A. implemented the algorithm; M.A., M.V., B.B., G.C. analysed the results; M.A. wrote the manuscript; M.V. and B.B edited the manuscript; B.B., J.K.Q. conceived the COPE study and collected the data. All the authors have read and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests. The corresponding author is responsible for submitting a competing interests statement on behalf of all authors of the paper.

Ethics approval, consent to participate

All the analyses performed in this work were carried out in accordance with relevant guidelines and regulations. All participants gave written informed consent at the recruitment wave to participate in the study and at each subsequent wave. The Research Ethics Committee for Camden and Islington provided ethical approval for the study (14/LO/2216). Approval was also granted by NHS Research and Development and the use of CPRD GOLD data was approved by the CPRD Independent Scientific Advisory Committee. The authors assert that all procedures contributing to this work comply with the relevant national and institutional committees' ethical standards on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-85089-2.

Correspondence and requests for materials should be addressed to M.V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.