International Journal of Chronic Obstructive Pulmonary Disease

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Development and Validation of an Electronic Health Record Algorithm to Predict the Presence of Chronic Obstructive Pulmonary Disease

Brian J Wells[1], Amit K Saha[2], Jill A Ohar[3]

[1]Department of Biostatistics and Data Science, Wake Forest University School of Medicine, Winston Salem, NC, USA; [2]Department of Anesthesiology, Wake Forest University School of Medicine, Winston Salem, NC, USA; [3]Department of Internal Medicine, Section on Pulmonary, Critical Care, Allergy and Immunologic Disease, Wake Forest University School of Medicine, Winston Salem, NC, USA

Correspondence: Brian J Wells, Department of Biostatistics and Data Science, Wake Forest University School of Medicine, Medical Center Blvd, Winston Salem, NC, 27185, USA, Tel +1 336 257-7128, Email bjwells@wakehealth.edu

International Journal of Chronic Obstructive Pulmonary Disease downloaded from https://www.dovepress.com/
For personal use only.

Purpose: Chronic obstructive pulmonary disease (COPD) is frequently diagnosed and treated based on clinical suspicion alone, without spirometric confirmation of expiratory airflow obstruction (AFO, defined by a forced expiratory volume in 1 second (FEV_1) to forced vital capacity (FVC) ratio of < 0.7). This can lead to overdiagnosis and unnecessary medication use, whereas underdiagnosis results in missed treatment opportunities. The Global Initiative for Chronic Obstructive Lung Disease (GOLD) recommends targeted case finding. This study aimed to develop and validate an automated Electronic Health Record (EHR) based algorithm to predict AFO and guide targeted spirometric testing.

Patients and Methods: Our analysis included 15,065 patients who underwent pulmonary function testing between 2016–2022. Patients were categorized as having AFO (n=4632) or not (n=10,433) based on spirometry. Patients < 45 years, with cystic fibrosis, alpha-1 antitrypsin deficiency, or prior spirometric evidence of obstruction were excluded. Logistic regression assessed 65 variables, retaining those that optimized model discrimination. The data were randomly split into training (n=10,546) and validation (n=4519) sets.

Results: Key predictors of AFO included older age, male sex, lower BMI, smoking history, prior COPD diagnosis, increased emergency department utilization, fewer outpatient visits, fewer chest X-rays, and higher cumulative beta-agonist prescriptions. The final model achieved an area under the receiver operating characteristic curve (AUC) of 0.82 (95% CI: 0.81–0.83) in the validation dataset and was well calibrated.

Conclusion: We developed an EHR-based algorithm that accurately predicts AFO using routinely collected structured data. This tool provides a practical method for identifying patients for targeted COPD case finding. Future efforts will focus on external validation and integration into clinical workflows, enabling automated identification and provider or patient notification to facilitate appropriate pulmonary function testing.

Plain Language Summary: Researchers at the Wake Forest University School of Medicine found patterns in computerized medical records that can help identify patients with undiagnosed chronic obstructive pulmonary disease (COPD). Finding patients with possible COPD allows for testing, earlier diagnosis, and treatment before the disease becomes severe. COPD is a group of lung diseases (such as emphysema) that make it difficult to breathe. Although many people with COPD have a history of smoking, there are other risk factors as well. It is estimated that about 60% of people with COPD do not know they have it. COPD can be diagnosed easily with a painless breathing test called spirometry, which only takes a few minutes in a doctor's office. This new tool can automatically find patients who should be tested for COPD without needing them to fill out symptom questionnaires. It uses statistics to combine information such as age, sex, weight, height, and smoking history with data on past prescriptions, chest x-rays, and emergency department visits. The researchers hope that, in the future, patients might be notified about their risk and have the test done even before seeing their doctor. Early diagnosis allows people to get vaccines that protect their lungs and to start medicines that can help prevent serious health problems. By using this tool, doctors can help more patients get the care they need sooner.

https://doi.org/10.2147/COPD.S546651
Received: 12 June 2025
Accepted: 16 December 2025
Published: 6 January 2026

© 2026 Wells et al. This work is published and licensed by Dove Medical Press Limited. The full terms of this license are available at https://www.dovepress.com/terms.php and incorporate the Creative Commons Attribution – Non Commercial (unported, v4.0) License (http://creativecommons.org/licenses/by-nc/4.0/). By accessing the work you hereby accept the Terms. Non-commercial uses of the work are permitted without any further permission from Dove Medical Press Limited, provided the work is properly attributed. For permission for commercial use of this work, please see paragraphs 4.2 and 5 of our Terms (https://www.dovepress.com/terms.php).

Keywords: risk assessment, chronic obstructive pulmonary disease, spirometry, screening, clinical informatics, electronic health records

Introduction

COPD represents a significant public health challenge characterized by a discrepancy between clinical recognition and the actual prevalence of disease. Approximately 7% of adults > 45 years of age report a history of COPD.[1] Data from the National Health and Nutrition Examination Survey revealed that approximately 60% of cases are undiagnosed, making the true prevalence close to 15%.[2] While not truly asymptomatic, patients often fail to report their symptoms, attributing them to both aging and smoking. Therefore, when a diagnosis is made, it is only after severely affecting a patient's quality of life when they have lost 50% of their lung function.[3] Underdiagnosis deprives patients of appropriate care, such as effective medications, vaccinations, and pulmonary rehabilitation, which have been shown to improve lung function and quality of life while reducing exacerbations and mortality.[4,5]

COPD is frequently diagnosed and treated based on clinical suspicion alone without a pulmonary function test (PFT). Confirming a diagnosis of COPD requires evidence of expiratory airflow obstruction (AFO) with spirometry, a standard part of a PFT. AFO is defined by a forced expiratory volume in 1 second (FEV1) to forced vital capacity (FVC) ratio of < 0.7. Attempting to detect COPD earlier through screening is controversial. Citing a lack of evidence for a benefit among "asymptomatic" individuals, the United States Preventive Services Task Force has recommended against spirometric screening for COPD in "asymptomatic" individuals.[6] The GOLD Report[7] recommends targeted case finding by screening high-risk individuals and suggests two assessment tools published by Lambe[8] and Haroon.[9] These tools are designed to detect early symptomatic disease with physician-administered patient questionnaires. The documentation of symptoms in the EHR are inconsistent and are typically recorded in unstructured text, limiting their use for automated, widespread screening. To date, many reports describe using AI technology, such as natural language processing, to aid in COPD diagnosis; however, the integration of these tools into Electronic Health Records (EHRs) remains challenging.[10–15]

Risk stratification tools relying entirely on existing, structured electronic health record (EHR) data would provide an opportunity to identify high risk patients for spirometry in a simple automated fashion. While Kotz et al developed such a tool, its utility is limited by its reliance on ICD codes to define the 10-year risk of COPD.[16] The limited accuracy of ICD codes in identifying true COPD has been well-documented. Most patients who carry diagnosis codes for COPD in their medical records do not have spirometric confirmation of AFO.[17]

The purpose of this study was to develop and evaluate an automated EHR-based prediction model to assist with targeted case finding with spirometry to improve the diagnosis of COPD. Specifically, we aimed to determine whether readily available structured EHR data could be used to identify patients at high risk for undiagnosed COPD, and to assess the predictive performance of such a model. By clearly defining and testing this approach, we sought to establish a foundation for future clinical interventions that could guide spirometry ordering and ultimately reduce underdiagnosis.

Materials and Methods

This study complies with the Declaration of Helsinki and was approved by the Atrium Health - Wake Forest University School of Medicine, Institutional Review Board (IRB#00090876). The board approved a waiver of informed consent due to the retrospective nature of the review. Privacy was ensured through strict confidentiality measures, including restricting access to data on a secure server only available to authorized members of the research team.

We have attempted to follow the TRIPOD+AI guidelines[18] for the development and reporting of a prediction model, and a checklist is provided in the supplemental material. Data used for algorithm building, recorded on or before the date of performance of the index outpatient PFT, were extracted from a common electronic health record system from the Wake Forest Baptist Medical Center in Winston Salem, NC, and Lexington Medical Center in Lexington, NC. The study cohort comprised 15,065 patients aged ≥ 40 years with a single spirometry (PFT) performed between 2016 and 2022; individuals with multiple tests were excluded. We excluded patients with known cystic fibrosis and alpha-1 antitrypsin deficiency because they represent unique clinical situations with more severe lung disease at a younger age.

We used logistic regression to predict evidence of obstruction on spirometry (FEV1 / FVC < 0.7) at the time of the first PFT result in our system. The fixed ratio definition of obstruction (FEV1 / FVC < 0.7) was used because of its simplicity and ease of use across multiple patient care facilities, allowing more universal use. The model output is a probability. We chose the first PFT since only a small percentage of individuals underwent multiple testing, which may represent a different population. Almost all PFTs were performed in the outpatient setting when patients were considered to be at their baseline functional status (Table 1). We included in-hospital-performed PFTs since they have been shown to accurately represent airflow patterns (obstructed, PRISM, normal).[19] We focused on the binary outcome of obstructed vs non-obstructed patients on the basis of the clinical relevance of this distinction that guides treatment. The screening flow chart of the participants in this study is shown in Figure 1.

The candidate independent variables used for predicting obstruction were limited to readily available data existing in structured fields in the EHR. Encounter diagnoses were limited to the prior year to capture conditions documented in the context of recent care episodes. In contrast, problem list entries were not time-limited, as their inclusion signifies ongoing and clinically relevant conditions. Healthcare utilization and medication data were available from any prior time. Logistic

Table 1 Demographics, Healthcare Coverage, Utilization & Comorbidities

Variable	Normal (n=9234)	PRISM (n=1199)	Airflow Obstruction (n= 4632)
Demographics			
Age in Years* (median [IQR])	63.00 [54.00, 71.00]	62.00 [54.00, 70.00]	67.00 [59.00, 74.00]
BMI* (median [IQR])	30.20 [26.10, 35.20]	33.30 [27.90, 39.70]	26.60 [22.50, 31.20]
Smoker* (Past or Present), n (%)	5804 (62.9)	819 (68.3)	4099 (88.5)
Sex*, n (%)			
Female	5052 (54.7)	756 (63.1)	1864 (40.2)
Male	4182 (45.3)	443 (36.9)	2768 (59.8)
Race, n (%)			
White	6815 (73.8)	877 (73.1)	3848 (83.1)
Black	1957 (21.2)	268 (22.4)	654 (14.1)
Other	462 (5.0)	54 (4.5)	130 (2.8)
Ethnicity, Hispanic, Latino or Spanish, n (%)	254 (2.8)	22 (1.8)	54 (1.2)
Zip Code Area Deprivation Index National Rank[†](median [IQR])	71.00 [59.00, 82.00]	73.00 [61.00, 84.00]	71.00 [60.00, 82.00]
Insurance, n (%)			
Commercial	3647 (39.5)	457 (38.1)	1182 (25.5)
Medicaid	745 (8.1)	124 (10.3)	510 (11.0)
Medicare	4153 (45.0)	535 (44.6)	2629 (56.8)
Other Government	210 (2.3)	24 (2.0)	136 (2.9)
Self-Pay	479 (5.2)	59 (4.9)	175 (3.8)
Clinical Characteristics, n (%)			
COPD Diagnosis* (%)	1154 (12.5)	307 (25.6)	2443 (52.7)
GERD Diagnosis (%)	596 (6.5)	61 (5.1)	230 (5.0)

(*Continued*)

https://doi.org/10.2147/COPD.S546651

Table 1 (Continued).

Variable	Normal (n=9234)	PRISM (n=1199)	Airflow Obstruction (n= 4632)
Obesity Diagnosis ‡ (%)	490 (5.3)	103 (8.6)	145 (3.1)
Acute Bronchitis Diagnosis (%)	19 (0.2)	6 (0.5)	15 (0.3)
Pneumonia Diagnosis (%)	68 (0.7)	11 (0.9)	40 (0.9)
Asthma Diagnosis (%)	128 (1.4)	46 (3.8)	61 (1.3)
Chronic Kidney Disease Diagnosis (%)	751 (8.1)	77 (6.4)	243 (5.2)
Diabetes Diagnosis (%)	1002 (10.9)	122 (10.2)	366 (7.9)
Congestive Heart Failure Diagnosis (%)	729 (7.9)	114 (9.5)	376 (8.1)
Patient Setting			
Inpatient (%)	1593 (17.3)	133 (11.1)	796 (17.2)
Outpatient (%)	7641 (82.7)	1066 (88.9)	3836 (82.2)
Months Since First Encounter (median [IQR])	162.00 [38.00, 288.00]	216.00 [74.00, 332.00]	177.00 [43.00, 301.00]
Number of Previous Encounter Lifetime (median [IQR])	76.00 [22.00, 194.00]	110.00 [31.00, 255.50]	66.00 [20.00, 162.00]
Number of Outpatient visits past year* # (median [IQR])	22.00 [9.00, 45.00]	26.00 [11.00, 53.00]	18.00 [8.00, 38.00]
Number of Inpatient Admission Past Year # (Mean [SD])	0.31 (0.82)	0.57 (1.25)	0.42 (0.94)
Number of ED visits past year* # (Mean [SD])	0.64 (1.75)	1.14 (2.30)	0.76 (1.63)
Number of Chest X-rays past year* # (Mean [SD])	0.23 (1.18)	0.19 (1.34)	0.14 (0.95)
Number of earlier β Agonist NEB Prescriptions Lifetime* (Mean [SD])	0.95 (3.86)	3.65 (8.96)	7.51 (27.67)

Notes: # Measured from the date of spirometry. *In the final prediction model. †The Area Deprivation Index (ADI) is a composite measure of neighborhood socioeconomic disadvantage for the United States that was calculated based on patient address. ADI values are presented in national percentile rankings from 0 (least disadvantaged) to 100 (most disadvantaged.) Kind, A. J. H., Jencks, S., Brock, J., Yu, M., Bartels, C., Ehlenbach, W., Greenberg, C., and Smith, M. (2014). Neighborhood Socioeconomic Disadvantage and 30-Day Rehospitalization. Annals of Internal Medicine, 161(11), 765–774. https://doi.org/10.7326/M13-2946. ‡Coded diagnosis of obesity.

regression using structured data was chosen as the statistical method on the basis of its simplicity, explainability, and ease of implementation. However, we also compared the accuracy of our logistic regression model with that of a random forest model to ensure that we were not sacrificing accuracy. The candidate variables were chosen on the basis of their theoretical relationship with obstruction, clinical relevance, and evidence in the scientific literature.[20] The domains for the candidate variables included demographics, vital statistics, laboratory measurements, medications, diagnosis codes, and healthcare utilization. The clinical characteristics included all the diagnosis codes from the encounter and the problem list available at the time of PFT. A complete list of the candidate variables is included in the supplemental material. Missing data were imputed using the multiple imputation by chained equations method via the mice package in R,[21] yielding a single dataset based on all candidate variables and the outcome. To assess real world application of the model, we did not impute missing data in the prospective cohort.

The data were randomly split into 70% for model building and 30% for model testing, ensuring that no patients used to create the statistical model were included in the test dataset used to evaluate model performance. A full logistic regression model was fit to the model build dataset using all 65 of the candidate variables. Continuous predictor variables were fit via restricted cubic splines to allow the model to account for nonlinear relationships. Splines were later removed from the model after variable selection if they were not statistically significant. This was done to reduce the number of

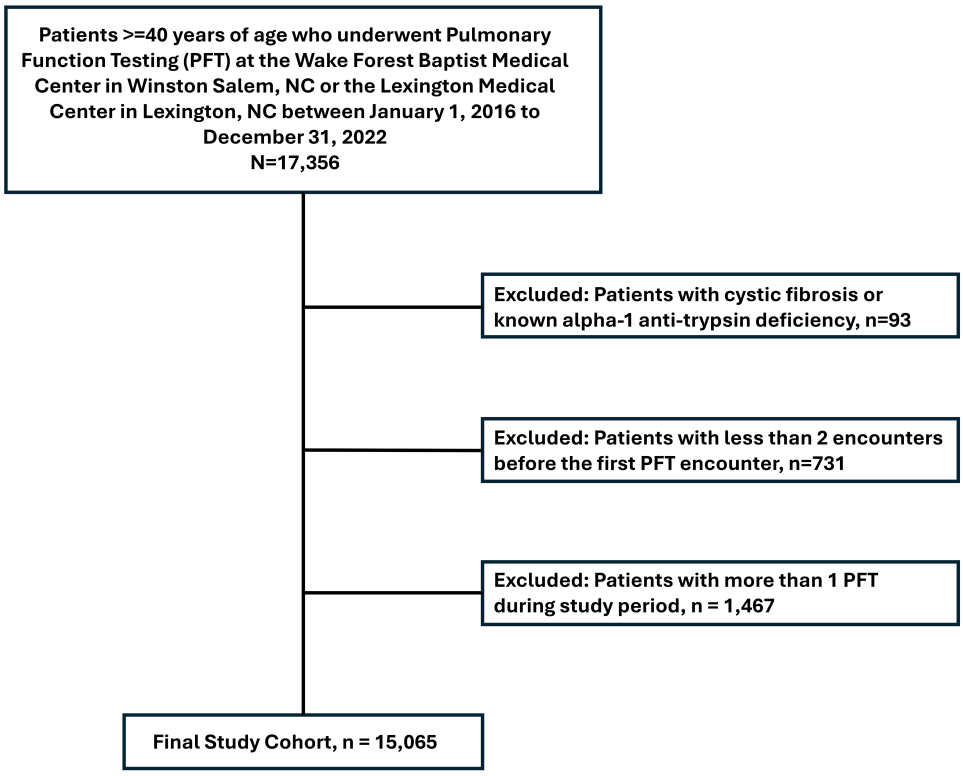

Figure 1 Study Flow Chart. The flow diagram illustrates patient selection, starting with all patients who underwent pulmonary function testing. After excluding patients with cystic fibrosis, alpha-1 anti-trypsin deficiency, less than 2 encounters, or more than one PFT, the final cohort included15,065 patients.

degrees of freedom in the model in an attempt to prevent overfitting and to allow smaller sample sizes for future model validation. The complete logistic regression model was reduced via Harrell's model approximation method.[22]

Performance metrics were calculated by applying the model created using the build dataset to the validation dataset. Discrimination was assessed via the area under the receiver operating characteristic curve (AUC). Calibration was assessed by plotting the predicted probability of obstruction with the actual incidence of obstruction. To assess clinical utility, we calculated classification metrics at different probability thresholds to guide screening spirometry and compared the model in its ability to identify cases of obstruction compared with screening the oldest smokers in our dataset. Sensitivity analyses were performed to assess model performance on the basis of whether spirometry was performed in an inpatient or outpatient setting.

To prospectively validate the tool, the risk equation was applied to data from the subsequent year, using identical eligibility criteria and outcome definitions. Variables were defined and processed as in model development and restricted to those retained in the final model. The dataset was used solely for performance assessment; no model building or imputation was performed, and patients with missing variables were excluded.

Results

Characteristics of the Study Cohort

A total of 15,065 subjects underwent pulmonary function testing and were included in the study. Among these patients, 4632 (30.7%) demonstrated obstruction. Among the subjects with a normal FEV1 / FVC, 1199 (15%) had preserved ratio impaired spirometry (PRISm), defined as an FEV1 /FVC $\geq$ 70%, but a reduction in the FEV1 and/or FVC < 80% of predicted. Table 1 shows the patient characteristics. The median age of the study cohort was 64 years (IQR 56–72). Among them, 7672 (51%) were male, and 11,540 were white (77%). The number of patients who were ever smokers was 10,215 (68%). Body mass index (BMI) was missing for 274 patients (1.82%), and smoking history or present information

was missing for 714 (4.7%) patients. The missing data were addressed via the multiple imputation using chained equation ("MICE") approach.

Table 1 shows the differences in health coverage between patients with and without obstruction. Patients with obstruction were older and therefore more likely to be on Medicare, which was the most common primary payer for all groups. Subjects with AFO had fewer outpatient visits and encounters with the healthcare system overall than did subjects with PRISM or who had normal spirometry (Table 1). However, they had more inpatient and ED visits than did subjects who had normal spirometry but fewer visits than did those with PRISM. Subjects with airflow obstruction more frequently were diagnosed with COPD, whereas subjects with normal spirometry and PRISM more commonly were diagnosed with GERD, obesity, diabetes and chronic kidney disease (Table 1). A history of acute bronchitis and pneumonia was not a differentiating factor among the 3 groups (Table 1), whereas obesity, asthma and congestive heart failure diagnoses tended to be more common in subjects with PRISM than in subjects with normal and obstructed spirometry.

Results of the Predictor Selection and Model Development

The study cohort was split into 2 datasets: the model build dataset (70%) and the test dataset (30%). The build and test datasets had 3264 and 1368 obstructed PFT patients, respectively. Multivariable logistic regression analysis after Harrell's model approximation resulted in a model with 9 variables identified as the most predictive of AFO included: older age, male sex, lower BMI, presence of a smoking history, past diagnosis of "COPD", increased numbers of ED visits, decreased outpatient visits, less frequent chest X-rays, and more lifetime beta agonist prescriptions. In the final model, the non-linear component for the number of chest x-rays was not statistically significant and the spline was removed. The overall model discrimination as variables were removed from the full model are shown in Table 2.

The receiver operating characteristic curve is shown in Figure 2. The AUC was 0.82 (95% CI 0.81–0.83) for the test data, and the model appeared to be well calibrated (Figure 3). The AUC was 0.83 (95% CI 0.81–0.84) for the spirometry conducted in outpatient settings (n= 12,543), whereas the AUC was 0.79 (95% CI 0.77–0.81) for the inpatient setting data (n= 2522). The classification metrics at different probability thresholds are shown in Table 3. To assess clinical utility, we estimated the number of cases of COPD that would be identified by testing 1000 individuals. Using the tool to identify the highest-risk patients, we identified 682 undiagnosed cases of COPD, whereas screening the 1000 oldest smokers identified only 449 patients.

Table 2 AUC Variation with Addition of Variables

Terms	AUC*
Coin Flip	0.5
Age	0.5991
Sex	0.6231
Ever Smoker	0.6967
Body Mass Index (BMI)	0.7458
COPD Diagnosis Code	0.8022
Number of Chest X-rays past year	0.8034
Number of ED visits past year	0.8038
Number of Outpatient visits past year	0.8063
Number of lifetime β Agonist Nebulizer prescriptions	0.8175

Note: *Area under the curve.

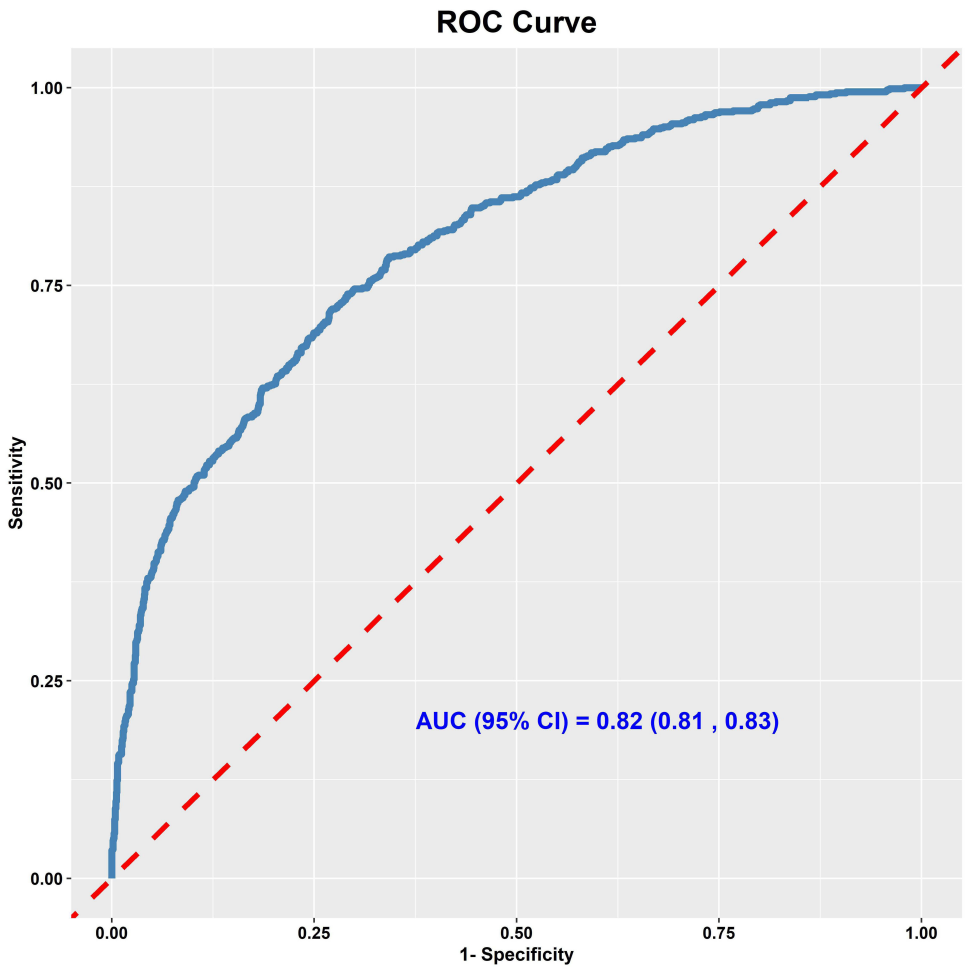

Figure 2 Receiver Operating Characteristic (ROC) Curve. The ROC curve illustrates the model's ability to differentiate between patients with and without the outcome. It displays the balance between sensitivity (true positive rate) and 1–specificity (false positive rate). The model's overall discriminative performance is reflected by the area under the curve (AUC) of 0.82.

Sensitivity Analysis

Prospective validation of the tool was performed on the basis of the data collected (n=1941) in the year after model development. The AUC for the risk equation for predicting obstruction in the prospective dataset was 0.79 (0.77–0.81).

Discussion

COPD is commonly diagnosed and treated in patients without the benefit of spirometric confirmation of airflow obstruction (AFO). This leads to potential overdiagnosis and underdiagnosis, both of which adversely impact outcomes and the cost of care.[17,23,24] Current clinical practice lacks an efficient, EHR-integrated tool for accurately predicting AFO, which is necessary for targeted COPD case finding. The aim of this study was to develop an EHR-based algorithm that uses only structured data to predict AFO that could be universally available to a variety of healthcare systems, even those without the advantage of high-tech computing systems. We were able to construct a model for targeted COPD case finding, with good accuracy and an AUC of 0.82, using 9 elements commonly available in the EHR. Our AUC compares favorably with the AUC of 0.74 obtained by Haroon et al.[9] However, comparing models in different datasets should be done with caution as the prevalence of the condition can impact model accuracy. We are unable to make a head-to-head comparison with the model by Haroon as the symptoms necessary to run their calculations are not available in our dataset. The clinical utility of our model is probably best highlighted by the hypothetical screening scenario described in the results illustrating the efficiency of our model in identifying new cases of COPD.

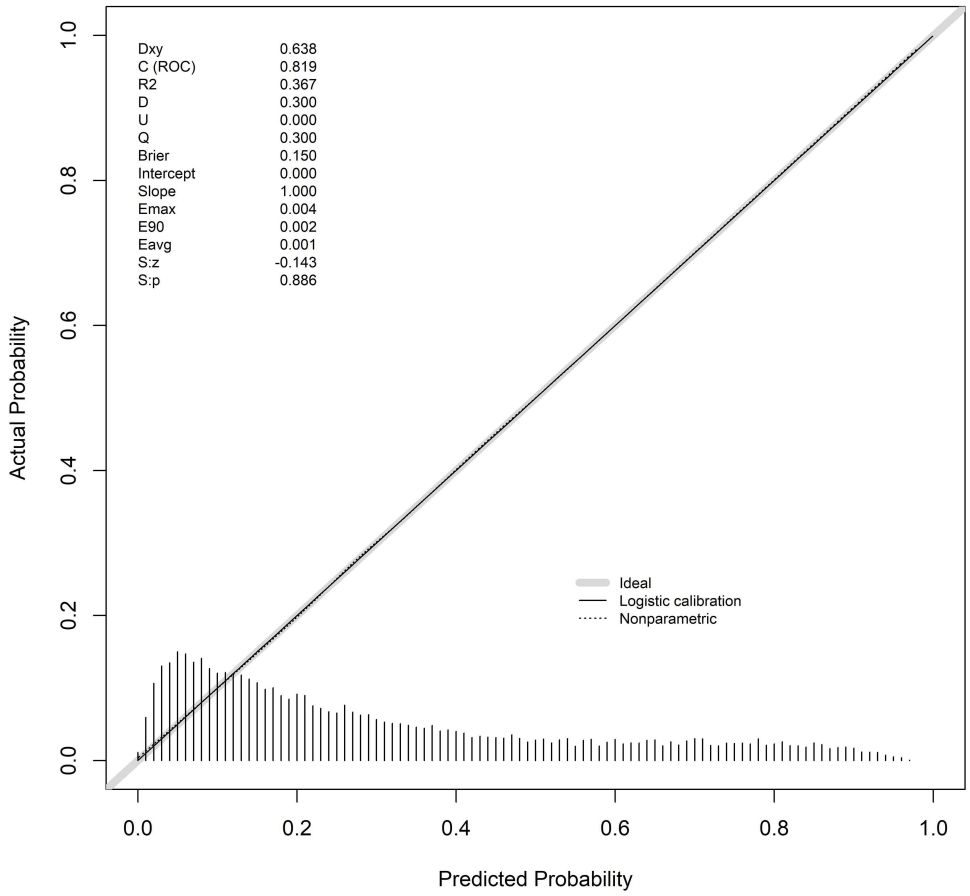

Figure 3 Calibration Curve. The plot compares predicted probabilities with observed event rates to assess calibration and is accompanied by a histogram of the predicted probabilities. The figure is annotated with key performance metrics; among them is the Brier score (0.15), a global measure that reflects both discrimination and calibration.

The supplemental material provides an equation for predicting the probability of airflow obstruction. The calculation could be implemented inside the EHR by local health information technology specialists at the organization or calculated from the EHR database outside of the clinical environment. The result of the model is a simple probability that could be used by a clinician or population health team to identify high-risk patients for screening. We also plan to evaluate the impact of a direct-to-patient alert for high-risk individuals who will trigger an automated order for spirometry for individuals who agree to testing. This strategy is similar to our work with targeted hemoglobin A1c testing.[25]

Data variability is always a concern when implementing prediction models in health systems outside of the system used to create the model. Data variability can arise from different patient characteristics, documentation practices, and the

Table 3 Classification at Different Probability Thresholds

Threshold for Intervention*	Sensitivity	Specificity	PPV	NPV
10%	0.95	0.33	0.38	0.94
25%	0.78	0.69	0.53	0.88
50%	0.5	0.9	0.7	0.81
75%	0.24	0.98	0.84	0.75

Notes: * Probability of airflow obstruction defined by a forced expiratory volume in 1 second (FEV_1) to forced vital capacity (FVC) ratio of <0.70.
Abbreviation: PPV, Positive predictive value; NPV, Negative predictive value.

completeness of EHR data across institutions. These factors can influence the performance of a model calibrated in one setting when applied to another. However, the simplicity of our model helps mitigate these challenges. The variables it uses are widely available and consistently recorded across diverse health systems, reducing reliance on site-specific documentation patterns. Furthermore, our statistical approach enables other institutions to recalibrate model coefficients using their own data, without resorting to complex methods like artificial neural networks that are often difficult to integrate into clinical workflows. This design lowers barriers to real-world implementation and supports robust performance across varied healthcare environments.

To implement the model, health systems need to calculate the values of the 9 variables, which can be done via readily available structured data. Calculation will require access to prescriptions and orders for chest X-rays, and decisions will need to be made about how to handle missing BMI values and smoking history. Given the small numbers, we do not calculate probabilities for individuals with missing information during implementation. We recommend local validation of the tool, and users may consider refitting or recalibration of the model to their data. The tool is intended only to guide targeted screening for case finding and is not intended to be used to diagnose COPD or to guide treatment decisions. Therefore, FDA approval is not being pursued.

Our work has several limitations. First and foremost, there is the potential for bias. Our data, extracted from 2 hospitals within our medical system, a university medical center and a community hospital, may not reflect the general population. Therefore, the algorithm will require validation in a large general practice population, which is currently a work in progress. Once validated, it can be deployed into the EHR to run in the background and produce alerts to both patients and clinicians that a diagnostic spirogram is needed. Although the tool performs well, the sample of patients was limited to those who already had PFT data, which may overestimate the tool's accuracy. An evaluation of the tool in an unbiased population of patients is warranted before contributing resources for implementing a targeted screening intervention.

Our algorithm was developed using data from two hospital-based practices within the same network. One was a community hospital, and the other was a university hospital. The data used to generate the algorithm may be population specific and therefore may not perform as well in other types of medical practices or regions of the country or world. The tool is dependent upon the data available in the EHR system and will likely underestimate risk in newer patients with fewer encounters as well as those patients who seek care outside the health system where the tool is being deployed. Notably, the data analyzed to create the algorithm spanned 2016–2022, and the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic may have affected medical practice and influenced the data.

Validated questionnaires have been used with the goal of targeted case finding but are burdensome compared with an EHR-based algorithm. The available validated questionnaires include Capture,[26] the COPD population screening tool,[27] the lung function questionnaire[28] and the International Primary Care Group questionnaire.[29] These tend to rely on several common elements: age; sex; smoking status, including pack years; prior diagnosis of COPD; and the presence of symptoms such as cough, sputum production, wheezing and dyspnea.[30] Leidy et al identified what they referred to as "best variables for COPD case identification" via random forest analysis.[31] These variables, which were used to develop Capture, included symptoms, impact on activity and days of work missed. The screening questionnaires tend to perform well, with areas under the AUC ranging from 0.79–0.88,[32,33] but a fair head-to-head comparison with our model is not possible, as we do not have screening questionnaires in our dataset. To overcome this limitation, a validation study recruiting volunteers from a general population that includes both spirometry and several validated widely used questionnaires is currently underway by our research group. However, questionnaires have not achieved great utilization in clinical practice, likely because of their limitations, which include the time required to administer or score, which can be translated into costs, as well as the fact that responses are subject to patient recall. Patients tend to underestimate their symptom burden;[34] therefore, the reliability of a questionnaire is only as good as patient recall. Finally, additional tests, such as micro spirometers or peak flow measurements, improve questionnaire performance[35] but also add burden. Our goal was to identify high-risk individuals for targeted testing.

There are numerous reports in the literature describing machine learning algorithms for COPD.[11,13,15,36–48] While some focus on the diagnosis or phenotyping of COPD, others concentrate on COPD / asthma diagnosis and differentiation.[13–15] Others predict the risk for an acute exacerbation,[46–48] healthcare resource utilization or death

resulting from an exacerbation.[39–45] Nearly all of these algorithms are based on both structured and unstructured data, necessitating the use of natural language processing. Some require the use of advanced diagnostics such as genotyping,[36] and many require spirometric data.[13,15,48]

The benefit of algorithms is that they are built using claims, CT or EHR data and do not require a questionnaire. Mapel et al created an algorithm for the identification of undiagnosed COPD using administrative claims data.[3] This algorithm differs from ours in several ways: 1) Medication groupers are based on proprietary data available through the American Hospital Formulary service, which limits model use; 2) there are 19 variables in their model, making it more difficult to use; 3) they do not provide the formula for calculating their model on other datasets, which would involve an additional step; and 4) since they use claims data, it may not work as well on our data.

Cheng et al[49] reported a model that predicts COPD via diagnosis codes. It requires a user interface for manual implementation. Details are lacking from the report, but it does not appear that the model is easy to use. A machine learning model for diagnosing COPD was reported by Spathis.[13] The model lacks practical application for COPD prediction in that it requires both symptoms and pulmonary function test results. Both Hussain[50] and Exarchos[51] reviewed AI techniques in COPD. Some tools use natural language processing, image analyses, and deep learning. They attempt to predict a wide array of outcomes, such as COPD severity, progression, mortality and AECOPD. These methods are difficult to implement, and none of the models are limited to structured EHR data to create a simple tool for predicting the probability of AFO.

A principle for the development of a prediction model is that it should be reflective of the disease pathophysiology. COPD is a disorder characterized by intermittent flares or exacerbations (AECOPDs) that become more common as the disease progresses.[7] Therefore, it is intuitive that variables associated with AECOPD, such as chest X-rays, short-acting beta-agonists, and health care visits, over the preceding year are, as we demonstrated, significant elements of the model. Other medications indicating possible exacerbations, such as antitussives, antibiotics, and glucocorticoids, did not improve model performance, possibly because they are less specific or have significant collinearity with beta-agonists. Despite the potential inaccuracy of smoking data in the EHR and the lack of total exposure in pack-years, smoking was still an important variable for model prediction.

Conclusion

We developed a simple algorithm for predicting AFO that, unlike existing tools, relies entirely on structured EHR data. Its simplicity, ease of use, and accuracy make it a practical tool for risk stratification to guide targeted spirometry in high-risk individuals. Next steps include validation and head-to-head comparison with several validated COPD questionnaires.

Abbreviations

COPD, Chronic Obstructive Pulmonary Disease; AFO, Airflow Obstruction; EHR, Electronic Health Records; ED, Emergency Department; AUC, Area Under the receiver operating characteristic Curve; BMI, Body Mass Index; ICS, Inhaled Corticosteroids; LAMA, Long-Acting Muscarinic Agents; LABA, Long-Acting Beta-Agonist; GOLD, Global Initiative for Chronic Obstructive Lung Disease; PFT, Pulmonary Function Testing; IRB, Institutional Review Board; AECOPD, Acute Exacerbation of Chronic Obstructive Pulmonary Disease; FEV1, Forced Expiratory Volume in 1 second; FVC, Forced Vital Capacity; PRISm - Preserved Ratio Impaired SpiroMetry; MICE, Multiple Imputation with Chained Equations; GERD, Gastroesophageal Reflux Disease; PDE-4 - Phosphodiesterase-4 Inhibitor; SABA, Short-Acting Beta-Agonist.

Acknowledgments

This project was funded by a Diagnostic Quality Improvement (DxQI) Seed Grant from the Society to Improve Diagnosis in Medicine. JO and BW conceived of this project and developed the protocol. AS performed the data procurement and analysis. All the authors participated in writing the manuscript.

Disclosure

BW: no conflicts of interest in this work.

AS: no conflicts of interest in this work.

JO: Consulting/Advisory boards: Chiesi Pharma, Astra Zeneca, Viatris, Mylan, Theravance, Verona, Genetech, Research Grants: Teva, Chiesi, Mylan, COPD Foundation.

References

1. Association AL. COPD trends brief. Available from: https://www.lung.org/research/trends-in-lung-disease/copd-trends-brief. Accessed March 24, 2024.
2. Ford ES, Mannino DM, Wheaton AG, Giles WH, Presley-Cantrell L, Croft JB. Trends in the prevalence of obstructive and restrictive lung function among adults in the united states: findings from the national health and nutrition examination surveys from 1988-1994 to 2007-2010. *Chest*. 2013;143(5):1395–1406. doi:10.1378/chest.12-1135
3. Mapel DW, Frost FJ, Hurley JS, et al. An algorithm for the identification of undiagnosed COPD cases using administrative claims data. *J Manag Care Pharm*. 2006;12(6):457–465.
4. Rabe Klaus F, Martinez Fernando J, Ferguson Gary T, et al. Triple inhaled therapy at two glucocorticoid doses in moderate-to-very-severe COPD. *N Engl J Med*. 2020;383(1):35–48. doi:10.1056/NEJMoa1916046
5. Lipson DA, Barnhart F, Brealey N, et al. Once-daily single-inhaler triple versus dual therapy in patients with COPD. *N Engl J Med*. 2018;378 (18):1671–1680. doi:10.1056/NEJMoa1713901
6. US Preventive Services Task Force, Mangione CM, Barry MJ, Nicholson WK. Screening for chronic obstructive pulmonary disease: us preventive services task force reaffirmation recommendation statement. *JAMA*. 2022;327(18):1806–1811. doi:10.1001/jama.2022.5692
7. 2023 GOLD report. Global initiative for chronic obstructive lung disease - GOLD. Available from: https://goldcopd.org/2023-gold-report-2/. Accessed March 24, 2024.
8. Lambe T, Adab P, Jordan RE, et al. Model-based evaluation of the long-term cost-effectiveness of systematic case-finding for COPD in primary care. *Thorax*. 2019;74(8):730–739. doi:10.1136/thoraxjnl-2018-212148
9. Haroon S, Adab P, Riley RD, Fitzmaurice D, Jordan RE. Predicting risk of undiagnosed COPD: development and validation of the TargetCOPD score. *Eur Respir J*. 2017;49(6):1602191. doi:10.1183/13993003.02191-2016
10. Yang X. Application and prospects of artificial intelligence technology in early screening of chronic obstructive pulmonary disease at primary healthcare institutions in China. *Int J Chronic Obstr*. 2024;19:1061–1067. doi:10.2147/COPD.S458935
11. Zafari H, Langlois S, Zulkernine F, Kosowan L, Singer A. AI in predicting COPD in the Canadian population. *Biosystems*. 2022;211:104585. doi:10.1016/j.biosystems.2021.104585
12. Shen X, Liu H. Using machine learning for early detection of chronic obstructive pulmonary disease: a narrative review. *Respir Res*. 2024;25 (1):336. doi:10.1186/s12931-024-02960-6
13. Spathis D, Vlamos P. Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. *Health Informatics J*. 2019;25 (3):811–827. doi:10.1177/1460458217723169
14. Kaplan A, Cao H, FitzGerald JM, et al. Artificial intelligence/machine learning in respiratory medicine and potential role in asthma and COPD diagnosis. *J Allergy Clin Immunol*. 2021;9(6):2255–2261. doi:10.1016/j.jaip.2021.02.014
15. Kocks JWH, Cao H, Holzhauer B, et al. Diagnostic performance of a machine learning algorithm (asthma/chronic obstructive pulmonary disease [COPD] differentiation classification) tool versus primary care physicians and pulmonologists in asthma, COPD, and asthma/COPD overlap. *J Allergy Clin Immunol*. 2023;11(5):1463–1474.e3. doi:10.1016/j.jaip.2023.01.017
16. Kotz D, Simpson CR, Viechtbauer W, van Schayck OC, Sheikh A. Development and validation of a model to predict the 10-year risk of general practitioner-recorded COPD. *NPJ Prim Care Respir Med*. 2014;24(1):14011. doi:10.1038/npjpcrm.2014.11
17. Gershon AS, Thiruchelvam D, Chapman KR, et al. Health services burden of undiagnosed and overdiagnosed COPD. *Chest*. 2018;153 (6):1336–1346. doi:10.1016/j.chest.2018.01.038
18. Collins GS, Moons KG, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024:385:e078378. doi: 10.1136/bmj-2023-078378
19. Loh CH, Genese FA, Kannan KK, Lovings TM, Peters SP, Ohar JA. Spirometry in hospitalized patients with acute exacerbation of COPD accurately predicts post discharge airflow obstruction. *J COPD F*. 2018;5(2):124–133. doi:10.15326/jcopdf.5.2.2017.0169
20. Cavaillès A, Brinchault-Rabin G, Dixmier A, et al. Comorbidities of COPD. *Eur Respir Rev*. 2013;22(130):454. doi:10.1183/09059180.00008612
21. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *J Stat Softw*. 2011;45(3):1–67.
22. Harrell FE, Lee KL, Mark DB. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15:361–387. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4
23. Rice RD, Han X, Wang X, Al-Jaghbeer MJ. COPD overdiagnosis and its effect on 30-day hospital readmission rates. *Respiratory Care*. 2021;66 (1):11–17. doi:10.4187/respcare.07536
24. MeiLan K H, Ye W, Wang D, et al. Bronchodilators in tobacco-exposed persons with symptoms and preserved lung function. *N Engl J Med*. 2022;387(13):1173–1184. doi:10.1056/NEJMoa2204752
25. Wells BJ, Lenoir KM, Diaz-Garelli JF, et al. Predicting current glycated hemoglobin values in adults: development of an algorithm from the electronic health record. *JMIR Med Inform*. 2018;6(4):e10780–e10780. doi:10.2196/10780
26. Martinez FJ, Mannino D, Leidy NK, et al. A new approach for identifying patients with undiagnosed chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2017;195(6):748–756. doi:10.1164/rccm.201603-0622OC
27. Martinez FJ, Raczek AE, Seifer FD, et al. Development and initial validation of a self-scored COPD population screener questionnaire (COPD-PS). *COPD*. 2008;5(2):85–95. doi:10.1080/15412550801940721
28. Yawn BP, Mapel DW, Mannino DM, et al. Development of the Lung Function Questionnaire (LFQ) to identify airflow obstruction. *Int J Chronic Obstr*. 2010;5:1–10. doi:10.2147/COPD.S7683
29. Price DB, Tinkelman DG, Halbert RJ, et al. Symptom-based questionnaire for identifying COPD in smokers. *Respiration*. 2006;73(3):285–295. doi:10.1159/000090142

30. van Schayck CP, Halbert RJ, Nordyke RJ, Isonaka S, Maroni J, Nonikov D. Comparison of existing symptom-based questionnaires for identifying COPD in the general practice setting. *Respirology*. 2005;10(3):323–333. doi:10.1111/j.1440-1843.2005.00720.x
31. Leidy NK, Malley KG, Steenrod AW, et al. Insight into best variables for COPD case identification: a random forests analysis. *Chronic Obstr Pulm Dis*. 2016;3(1):406–418. doi:10.15326/jcopdf.3.1.2015.0144
32. Spyratos D, Haidich AB, Chloros D, Michalopoulou D, Sichletidis L. Comparison of three screening questionnaires for chronic obstructive pulmonary disease in the primary care. *Respiration*. 2017;93(2):83–89. doi:10.1159/000453586
33. Wang JM, Han MK, Labaki WW. Chronic obstructive pulmonary disease risk assessment tools: is one better than the others? *Curr Opin Pulm Med*. 2022;28(2):99–108. doi:10.1097/MCP.0000000000000833
34. Kessler R, Partridge MR, Miravitlles M, et al. Symptom variability in patients with severe COPD: a pan-European cross-sectional study. *Eur Respir J*. 2011;37(2):264–272. doi:10.1183/09031936.00051110
35. Schnieders E, Ünal E, Winkler V, et al. Performance of alternative COPD case-finding tools: a systematic review and meta-analysis. *Eur Respir Rev*. 2021;30(160):200350. doi:10.1183/16000617.0350-2020
36. Ma X, Wu Y, Zhang L, et al. Comparison and development of machine learning tools for the prediction of chronic obstructive pulmonary disease in the Chinese population. *J Transl Med*. 2020;18(1):146. doi:10.1186/s12967-020-02312-0
37. Er O, Temurtas F. A study on chronic obstructive pulmonary disease diagnosis using multilayer neural networks. *J Med Syst*. 2008;32(5):429–432. doi:10.1007/s10916-008-9148-6
38. Zhao Q, Li J, Zhao L, Zhu Z. Knowledge guided feature aggregation for the prediction of chronic obstructive pulmonary disease with Chinese EMRs. *IEEE/ACM Trans Comput Biol Bioinform*. 2023;20(6):3343–3352. doi:10.1109/TCBB.2022.3198798
39. Khatri KL, Tamil LS. Early detection of peak demand days of chronic respiratory diseases emergency department visits using artificial neural networks. *IEEE J Biomed Health Inform*. 2018;22(1):285–290. doi:10.1109/JBHI.2017.2698418
40. Min X, Yu B, Wang F. Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: a case study on COPD. *Sci Rep*. 2019;9(1):2362. doi:10.1038/s41598-019-39071-y
41. Swaminathan S, Qirko K, Smith T, et al. A machine learning approach to triaging patients with chronic obstructive pulmonary disease. *PLoS One*. 2017;12(11):e0188532. doi:10.1371/journal.pone.0188532
42. Agarwal A, Baechle C, Behara R, Zhu X. A natural language processing framework for assessing hospital readmissions for patients with COPD. *IEEE J Biomed Health Inform*. 2018;22(2):588–596. doi:10.1109/JBHI.2017.2684121
43. Goto T, Camargo CA, Faridi MK, Yun BJ, Hasegawa K. Machine learning approaches for predicting disposition of asthma and COPD exacerbations in the ED. *Am J Emergency Med*. 2018;36(9):1650–1654. doi:10.1016/j.ajem.2018.06.062
44. Goto T, Jo T, Matsui H, Fushimi K, Hayashi H, Yasunaga H. Machine learning-based prediction models for 30-day readmission after hospitalization for chronic obstructive pulmonary disease. *COPD*. 2019;16(5–6):338–343. doi:10.1080/15412555.2019.1688278
45. Peng J, Chen C, Zhou M, Xie X, Zhou Y, Luo CH. A machine-learning approach to forecast aggravation risk in patients with acute exacerbation of chronic obstructive pulmonary disease with clinical indicators. *Sci Rep*. 2020;10(1):3118. doi:10.1038/s41598-020-60042-1
46. Wang C, Chen X, Du L, Zhan Q, Yang T, Fang Z. Comparison of machine learning algorithms for the identification of acute exacerbations in chronic obstructive pulmonary disease. *Comput. Methods Programs Biomed*. 2020;188:105267. doi:10.1016/j.cmpb.2019.105267
47. Nunavath V, Goodwin M, Fidje JT, Moe CE. Deep neural networks for prediction of exacerbations of patients with chronic obstructive pulmonary disease. In: Pimenidis E, Jayne C, editors. *Engineering Applications of Neural Networks*. Springer International Publishing; 2018:217–228.
48. Ying J, Dutta J, Guo N, et al. Classification of exacerbation frequency in the COPDGene cohort using deep learning with deep belief networks. *IEEE J Biomed Health Inform*. 2020;24(6):1805–1813. doi:10.1109/JBHI.2016.2642944
49. Cheng YT, Lin YF, Chiang KH, Tseng VS. Mining sequential risk patterns from large-scale clinical databases for early assessment of chronic diseases: a case study on chronic obstructive pulmonary disease. *IEEE J Biomed Health Inform*. 2017;21(2):303–311. doi:10.1109/JBHI.2017.2657802
50. Hussain A, Marlowe S, Ali M, et al. A systematic review of artificial intelligence applications in the management of lung disorders. *Cureus*. 2024;16(1):e51581. doi:10.7759/cureus.51581
51. Exarchos KP, Aggelopoulou A, Oikonomou A, et al. Review of artificial intelligence techniques in chronic obstructive lung disease. *IEEE J Biomed Health Inform*. 2022;26(5):2331–2338. doi:10.1109/JBHI.2021.3135838

International Journal of Chronic Obstructive Pulmonary Disease

Dovepress
Taylor & Francis Group

Publish your work in this journal

The International Journal of COPD is an international, peer-reviewed journal of therapeutics and pharmacology focusing on concise rapid reporting of clinical studies and reviews in COPD. Special focus is given to the pathophysiological processes underlying the disease, intervention programs, patient focused education, and self management protocols. This journal is indexed on PubMed Central, MedLine and CAS. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit http://www.dovepress.com/testimonials.php to read real quotes from published authors.

Submit your manuscript here: https://www.dovepress.com/international-journal-of-chronic-obstructive-pulmonary-disease-journal