Developing a Machine Learning Model to Predict Severe Chronic Obstructive Pulmonary Disease Exacerbations: Retrospective Cohort Study

Siyang Zeng[1], MS; **Mehrdad Arjomandi**[2,3], MD; **Yao Tong**[1], MS; **Zachary C. Liao**[1], MD, MPH; **Gang Luo**[1], PhD
[1]Department of Biomedical Informatics and Medical Education, University of Washington, UW Medicine South Lake Union, 850 Republican Street, Building C, Box 358047, Seattle, WA 98195, USA
[2]Medical Service, San Francisco Veterans Affairs Medical Center, 4150 Clement Street, San Francisco, CA 94121, USA
[3]Department of Medicine, University of California, 505 Parnassus Avenue, San Francisco, CA 94143, USA
szeng998@uw.edu, mehrdad.arjomandi@ucsf.edu, yaotong@uw.edu, zachcl@uw.edu, luogang@uw.edu

Corresponding author:
Gang Luo, PhD
Department of Biomedical Informatics and Medical Education, University of Washington, UW Medicine South Lake Union, 850 Republican Street, Building C, Box 358047, Seattle, WA 98195, USA
Phone: 1-206-221-4596
Fax: 1-206-221-2671
Email: luogang@uw.edu

Abstract

Background: Chronic obstructive pulmonary disease (COPD) poses a large burden on healthcare. Severe COPD exacerbations require emergency department visits or inpatient stays, often cause irreversible decline in lung function and health status, and account for 90.3% of the total medical cost related to COPD. Many severe COPD exacerbations are deemed preventable with appropriate outpatient care. Current models for predicting severe COPD exacerbations lack accuracy, making it difficult to effectively target high-risk patients for preventive care management to reduce severe COPD exacerbations and improve outcomes.

Objective: To develop a more accurate model to predict severe COPD exacerbations.

Methods: We examined all patients with COPD who visited the University of Washington Medicine (UWM) facilities between 2011 and 2019 and identified 278 candidate features. By doing secondary analysis on 43,576 UWM data instances from 2011 to 2019, we created a machine learning model to predict severe COPD exacerbations in the next year for patients with COPD.

Results: The final model had an area under the receiver operating characteristic curve of 0.866. When using the top 10.00% (752/7,529) of patients with the largest predicted risk to set the cutoff threshold for binary classification, the model gained an accuracy of 90.33% (6,801/7,529), a sensitivity of 56.6% (103/182), and a specificity of 91.17% (6,698/7,347).

Conclusions: Our model provided more accurate prediction of severe COPD exacerbations in the next year compared with prior published models. After further improvement of its performance measures (e.g., by adding features extracted from clinical notes), our model could be used in a decision support tool to guide identification of high-risk patients with COPD for care management to improve outcomes.

International Registered Report Identifier (IRRID): PRR2-10.2196/13783

Keywords: Chronic obstructive pulmonary disease; machine learning; forecasting; symptom exacerbation; patient care management

Introduction

Background

In the United States, chronic obstructive pulmonary disease (COPD) affects 6.5% of adults [1] and is the fourth leading cause of death excluding coronavirus disease 2019 (COVID-19) [2]. Each year, COPD causes 1.5 million emergency department (ED) visits, 0.7 million inpatient stays, and US $32.1 billion in total medical cost [1]. Severe COPD exacerbations are those requiring ED visits or inpatient stays [3], account for 90.3% of the total medical cost related to COPD [4], and often cause irreversible decline in lung function and health status [5-10]. Many severe COPD exacerbations (e.g., 47% of inpatient stays for COPD) are deemed preventable with appropriate outpatient care [3,11], as COPD is an ambulatory care sensitive condition [12]. A commonly used method to reduce severe COPD exacerbations is to place high-risk patients in a care management program for preventive care [13-15]. High-risk patients can be identified prospectively using a predictive model [16]. Once a patient enters the care management program, a care manager will periodically contact the patient for health status assessment and to help coordinate health and related services. This method is adopted by many health plans, such as those in 9 of 12 metropolitan communities [13], and many healthcare systems. Successful care management can reduce up to 27% of ED visits [14] and 40% of inpatient stays [15] in patients with COPD.

However, due to limitations of resources and service capacity, only ≤3% of patients could enter a care management program [17]. Its effectiveness is upper bounded by these patients' risk levels, which are determined by how accurate the employed predictive model is. Neither the stage of COPD nor having prior severe COPD exacerbations alone can predict a patient's risk level for future severe COPD exacerbations well [18,19]. Previously, researchers had built several models to predict severe COPD exacerbations in patients with COPD [20-53]. These models are inaccurate and suboptimal for use in care management, as they missed over 50% of the patients who will experience severe COPD exacerbations in the future, incorrectly projected many other patients to experience severe COPD exacerbations [20-22,53], used data unavailable in routine clinical practice [23-31,33,34,36,42-50,52], or were designed for subjects who have different characteristics from typical patients with COPD [25-34]. Also, most of these models predicted only inpatient stays for COPD. To better guide the use of care management, we need to predict both ED visits and inpatient stays for COPD, which only 2 of these models [34,36] do. In practice, once a model is deployed for care management, the prediction errors produced by the model would lead to degraded patient outcomes and unnecessary healthcare costs. Due to the large number of patients with COPD, even a small improvement in model accuracy coupled with appropriate preventive interventions could help improve outcomes and avoid many ED visits and inpatient stays for COPD every year.

Objectives

This study aimed to develop a more accurate model to predict severe COPD exacerbations in the next year in patients with COPD. To be suitable for use in care management, the model should use data available in routine clinical practice and target all patients with COPD.

Methods

Ethics approval and study design

University of Washington Medicine (UWM)'s institutional review board approved this secondary analysis study on administrative and clinical data.

Patient population

In Washington State, the UWM is the largest academic healthcare system. The UWM enterprise data warehouse includes administrative and clinical data from 3 hospitals and 12 clinics. The patient cohort consisted of the patients with COPD who visited any of those facilities between 2011 and 2019. Using our prior method for identifying patients with COPD [54] that was adapted from the literature [55-58], we regarded a patient to have COPD if the patient was aged 40 or older and met 1 or more of the 4 criteria listed in Table 1. When computing the data instances in any year, we excluded the patients who had no encounter at the UWM or died during that year. No other exclusion criterion was used.

Table 1. The 4 criteria used for identifying patients with COPD.

Sequence number	Description of the criterion
1	An outpatient visit diagnosis code of COPD (International Classification of Diseases, Ninth Revision (ICD-9): 491.22, 491.21, 491.9, 491.8, 493.2x, 492.8, 496; International Classification of Diseases, Tenth Revision (ICD-10): J42, J41.8, J44.*, J43.*) followed by ≥1 prescription of long-acting muscarinic antagonist (aclidinium, glycopyrrolate, tiotropium, and umeclidinium) within 6 months
2	≥1 ED or ≥2 outpatient visit diagnosis codes of COPD (ICD-9: 491.22, 491.21, 491.9, 491.8, 493.2x, 492.8, 496; ICD-10: J42, J41.8, J44.*, J43.*)
3	≥1 inpatient stay discharge having a principal diagnosis code of COPD (ICD-9: 491.22, 491.21, 491.9, 491.8, 493.2x, 492.8, 496; ICD-10: J42, J41.8, J44.*, J43.*)
4	≥1 inpatient stay discharge having a principal diagnosis code of respiratory failure (ICD-9: 518.82, 518.81, 799.1, 518.84; ICD-10: J96.0*, J80, J96.9*, J96.2*, R09.2) and a secondary diagnosis code of acute COPD exacerbation (ICD-9: 491.22, 491.21, 493.22, 493.21; ICD-10: J44.1, J44.0)

Prediction target (also known as the outcome or the dependent variable)

Given a patient with COPD who had ≥1 encounter at the UWM in a specific year (the index year), we used the patient's data up to the last day of the year to predict the outcome of whether the patient would experience any severe COPD exacerbation, i.e., any ED visit or inpatient stay with a principal diagnosis of COPD (ICD-9: 491.22, 491.21, 491.9, 491.8, 493.2x, 492.8, 496; ICD-10: J42, J41.8, J44.*, J43.*), in the next year (Figure 1).

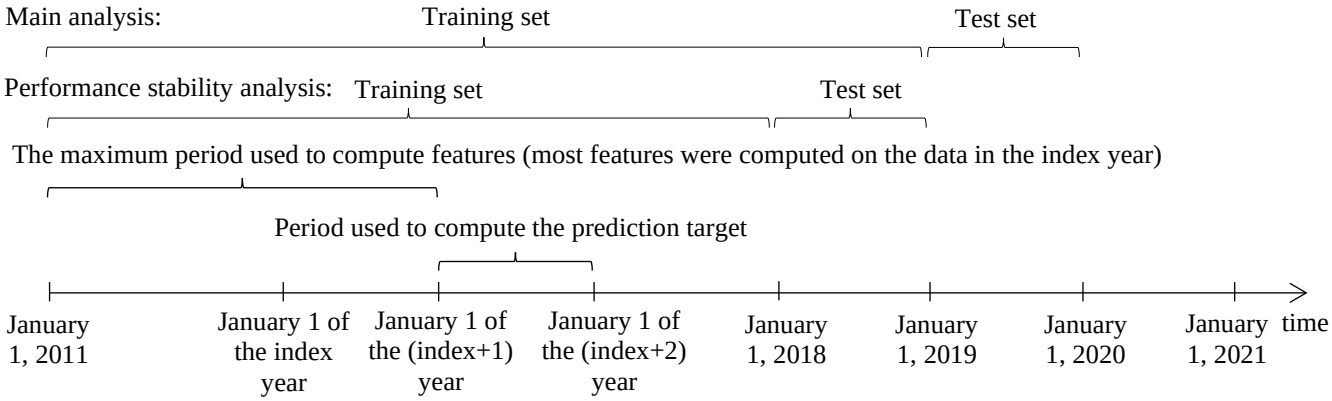

Figure 1. The time periods used to partition the training and test sets and the time periods used to compute the prediction target and the features for a (patient, index year) pair.

Data set

We obtained a structured data set from the UWM enterprise data warehouse. This data set included administrative and clinical data relating to the patient cohort's encounters at the 3 hospitals and 12 clinics of the UWM between 2011 and 2020.

Features (also known as independent variables)

To improve model accuracy, we examined an extensive set of candidate features computed on the structured attributes in the data set. Table 1 of the Appendix shows these 278 candidate features coming from 4 sources: the known risk factors for COPD exacerbations [3,18,28,30,50,59-72], the features used in the prior models to predict severe COPD exacerbations [20-53], the features that the clinician ZL in our team suggested, and the features used in our prior models to predict asthma hospital encounters [73,74]. Asthma shares many similarities with COPD. Throughout this paper, whenever we mention the number of a given type of items (e.g., medications) without using the word "distinct," we count multiplicity.

Each input data instance to the predictive model contained 278 features, corresponded to a distinct (patient, index year) pair, and was used to predict the outcome of the patient in the next year. For this pair, the patient's age was computed based on the end of the index year. The patient's primary care provider (PCP) was computed as the last recorded PCP of the patient by the end of the index year. The percentage of the PCP's patients with COPD in the pre-index year having severe COPD exacerbations in the index year was computed on the data in the pre-index and index years. Using the data from 2011 to the index year, we computed 26 features: the number of years from the first encounter related to COPD in the data set, the type of the first encounter related to COPD in the data set, 7 allergy features, and 17 features related to the problem list. The other 251 features were computed on the data in the index year.

Data analysis

Data preparation

Using the data preparation approach employed in our papers [73,74], we identified the biologically implausible values, replaced them with null values, and normalized the data. Since outcomes came from the next year, the data set had 9 years of effective data (2011-2019) over a time span of 10 years (2011-2020). To reflect future model use in clinical practice and to evaluate the impact of the COVID-19 pandemic on patient outcomes and model performance, we conducted 2 analyses:

1) Main analysis: We used the 2011-2018 data instances as the training set to train models and the 2019 data instances as the test set to assess model performance.
2) Performance stability analysis: We used the 2011-2017 data instances as the training set to train models and the 2018 data instances as the test set to assess model performance.

Classification algorithms

We created machine learning classification models using Waikato Environment for Knowledge Analysis (Weka) Version 3.9 [75]. Weka is a major open-source software package for machine learning and data mining. It integrates many commonly used machine learning algorithms and feature selection techniques. We examined the 39 classification algorithms supported by Weka and listed in the online appendix of our paper [73], as well as extreme gradient boosting (XGBoost) [76] implemented in the XGBoost4J package [77]. XGBoost is a classification algorithm using an ensemble of decision trees. As XGBoost only takes numerical features, we converted categorical features to binary features via one-hot encoding. In the main analysis, we used the training set and our formerly published automatic machine learning model selection method [78] to automate the selection of the classification algorithm, feature selection technique, data balancing method to deal with imbalanced data, and hyper-parameter values among all applicable ones. Compared with the Auto-WEKA automatic machine learning model selection method [79], our method achieved an average of 11% reduction in model error rate and a 28-fold reduction in search time. In the performance stability analysis, we used the same classification algorithm, feature selection technique, and hyper-parameter values as those used in the final model of the main analysis.

Performance metrics

Table 2. The confusion matrix.

Outcome class	Severe COPD exacerbations in the next year	No severe COPD exacerbation in the next year
Predicted severe COPD exacerbations in the next year	True positive (TP)	False positive (FP)
Predicted no severe COPD exacerbation in the next year	False negative (FN)	True negative (TN)

As shown in the formulas below, performance of the models was evaluated with respect to the following metrics: accuracy, sensitivity also known as recall, specificity, positive predictive value (PPV) also known as precision, negative predictive value (NPV), and area under the receiver operating characteristic curve (AUC).

accuracy = (TP+TN)/(TP+TN+FP+FN) (Table 2),
sensitivity = TP/(TP+FN),
specificity = TN/(TN+FP),
positive predictive value = TP/(TP+FP),
negative predictive value = TN/(TN+FN).

We computed the 95% confidence intervals (CIs) of the performance measures using the bootstrapping method [80]. We obtained 1,000 bootstrap samples from the test set and computed the model's performance measures based on each bootstrap sample. This produced 1,000 values for each performance metric. Their 2.5th and 97.5th percentiles provided the 95% CI of the corresponding performance measures. To depict the tradeoff between sensitivity and specificity, we drew the receiver operating characteristic curve.

Results

Distributions of data instances and bad outcomes

The number of data instances increased over time. The proportion of data instances linked to bad outcomes remained relatively stable over time. The only exception was the sudden drop from 5.21% (369/7,089) in 2018 to 2.42% (182/7,529) in 2019 (Table 3), which resulted from the large drop of ED visits and inpatient stays for COPD in 2020 caused by COVID-19 [81]. In the main analysis, 5.66% (2,040/36,047) of data instances in the training set and 2.42% (182/7,529) of data instances in the test set were linked to severe COPD exacerbations in the next year. In the performance stability analysis, 5.77% (1,671/28,958) of data instances in the training set and 5.21% (369/7,089) of data instances in the test set were linked to severe COPD exacerbations in the next year.

Table 3. The distributions of data instances and bad outcomes over time.

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019
Number of data instances	1,848	2,725	3,204	4,009	4,875	5,793	6,504	7,089	7,529
Number of data instances linked to severe COPD exacerbations in the next year, *n* (%)	128 (6.93)	176 (6.46)	183 (5.71)	223 (5.56)	272 (5.58)	351 (6.06)	338 (5.20)	369 (5.21)	182 (2.42)

Patient characteristics

Each (patient, index year) pair matched a data instance. For both the training set and the test set of the main analysis, when comparing the patient characteristic distributions between the data instances linked to severe COPD exacerbations in the next year and those linked to no severe COPD exacerbation in the next year, *P* values were computed using the χ^2 2-sample test and the Cochran-Armitage trend test [82] for categorical and numerical characteristics, respectively (Tables 4 and 5). *P* values <.05 are italicized and signify statistically significant differences in the patient characteristic distributions.

Table 4. The patient characteristics of the data instances in the training set of the main analysis.

Patient characteristic	Data instances (*N*=36,047), *n* (%)	Data instances linked to severe COPD exacerbations in the next year (*N*=2,040), *n* (%)	Data instances linked to no severe COPD exacerbation in the next year (*N*=34,007), *n* (%)	*P* value
Age				
40 to 65	18,793 (52.13)	1,219 (59.75)	17,574 (51.68)	*<0.001*
65+	17,254 (47.87)	821 (40.25)	16,433 (48.32)	
Sex				
Female	15,414 (42.76)	749 (36.72)	14,665 (43.12)	*<0.001*
Male	20,633 (57.24)	1,291 (63.28)	19,342 (56.88)	
Race				
American Indian or Alaska Native	713 (1.98)	26 (1.27)	687 (2.02)	*<0.001*
Asian	2,092 (5.80)	144 (7.06)	1,948 (5.73)	
Black or African American	4,795 (13.30)	524 (25.69)	4,271 (12.56)	
Native Hawaiian or other Pacific Islander	184 (0.51)	8 (0.39)	176 (0.52)	
White	27,447 (76.14)	1,330 (65.20)	26,117 (76.80)	
Other, unknown, or not reported	816 (2.27)	8 (0.39)	808 (2.37)	
Ethnicity				
Hispanic	857 (2.38)	53 (2.60)	804 (2.36)	*<0.001*
Non-Hispanic	32,585 (90.39)	1,941 (95.15)	30,644 (90.11)	

	Unknown or not reported	2,605 (7.23)	46 (2.25)	2,559 (7.53)	
Smoking status				
	Current smoker	16,952 (47.03)	1,089 (53.38)	15,863 (46.65)	*<0.001*
	Former smoker	7,367 (20.44)	345 (16.91)	7,022 (20.65)	
	Never smoker or unknown	11,728 (32.53)	606 (29.71)	11,122 (32.70)	
Insurance				
	Private	17,513 (48.58)	834 (40.88)	16,679 (49.05)	*<0.001*
	Public	29,598 (82.11)	1,767 (86.62)	27,831 (81.84)	*<0.001*
	Self-paid or charity	1,994 (5.53)	229 (11.23)	1,765 (5.19)	*<0.001*
Number of years from the first encounter related to COPD in the data set				
	≤3	30,315 (84.10)	1,566 (76.76)	28,749 (84.54)	*<0.001*
	>3	5,732 (15.90)	474 (23.24)	5,258 (15.46)	
COPD medication prescription				
	Inhaled corticosteroid (ICS)	13,327 (36.97)	1,119 (54.85)	12,208 (35.90)	*<0.001*
	Short-acting muscarinic antagonist (SAMA)	9,608 (26.65)	1,042 (51.08)	8,566 (25.19)	*<0.001*
	Short-acting beta-2 agonist (SABA)	22,549 (62.55)	1,684 (82.55)	20,865 (61.36)	*<0.001*
	SABA and SAMA combination	7,174 (19.90)	810 (39.71)	6,364 (18.71)	*<0.001*
	Long-acting muscarinic antagonist (LAMA)	10,243 (28.42)	1,001 (49.07)	9,242 (27.18)	*<0.001*
	Long-acting beta-2 agonist (LABA)	8,904 (24.70)	842 (41.27)	8,062 (23.71)	*<0.001*
	LABA and LAMA combination	426 (1.18)	40 (1.96)	386 (1.14)	*0.001*
	ICS and LABA combination	8,326 (23.10)	782 (38.33)	7,544 (22.18)	*<0.001*
	ICS, LABA, and LAMA combination	16 (0.04)	0 (0.00)	16 (0.05)	0.66
	Phosphodiesterase-4 inhibitor	94 (0.26)	10 (0.49)	84 (0.25)	0.06
	Systemic corticosteroid	11,293 (31.33)	1,144 (56.08)	10,149 (29.84)	*<0.001*
Comorbidity				
	Allergic rhinitis	2,445 (6.78)	174 (8.53)	2,271 (6.68)	*0.001*
	Anxiety or depression	10,786 (29.92)	725 (35.54)	10,061 (29.59)	*<0.001*
	Asthma	4,794 (13.30)	417 (20.44)	4,377 (12.87)	*<0.001*
	Congestive heart failure	6,063 (16.82)	495 (24.26)	5,568 (16.37)	*<0.001*
	Diabetes	7,623 (21.15)	446 (21.86)	7,177 (21.10)	0.43
	Eczema	1,558 (4.32)	98 (4.80)	1,460 (4.29)	0.30
	Gastroesophageal reflux	7,162 (19.87)	507 (24.85)	6,655 (19.57)	*<0.001*
	Hypertension	18,361 (50.94)	1,150 (56.37)	17,211 (50.61)	*<0.001*
	Ischemic heart disease	7,420 (20.58)	486 (23.82)	6,934 (20.39)	*<0.001*
	Lung cancer	794 (2.20)	52 (2.55)	742 (2.18)	0.31
	Obesity	3,487 (9.67)	255 (12.50)	3,232 (9.50)	*<0.001*
	Sinusitis	1,382 (3.83)	83 (4.07)	1,299 (3.82)	0.61
	Sleep apnea	3,179 (8.82)	253 (12.40)	2,926 (8.60)	*<0.001*

Table 5. The patient characteristics of the data instances in the test set of the main analysis.

Patient characteristic	Data instances (*N*=7,529), *n* (%)	Data instances linked to severe COPD exacerbations in the next year (*N*=182), *n* (%)	Data instances linked to no severe COPD exacerbation in the next year (*N*=7,347), *n* (%)	*P* value
Age				
40 to 65	3,442 (45.72)	118 (64.8)	3,324 (45.24)	*<0.001*
65+	4,087 (54.28)	64 (35.2)	4,023 (54.76)	
Sex				
Female	3,289 (43.68)	47 (25.8)	3,242 (44.13)	*<0.001*
Male	4,240 (56.32)	135 (74.2)	4,105 (55.87)	
Race				
American Indian or Alaska Native	156 (2.07)	5 (2.7)	151 (2.06)	*<0.001*
Asian	439 (5.83)	7 (3.9)	432 (5.88)	
Black or African American	896 (11.90)	57 (31.3)	839 (11.42)	

	Native Hawaiian or other Pacific Islander	53 (0.71)	2 (1.1)	51 (0.69)	
	White	5,793 (76.94)	111 (61.0)	5,682 (77.34)	
	Other, unknown, or not reported	192 (2.55)	0 (0.0)	192 (2.61)	
Ethnicity					
	Hispanic	188 (2.50)	3 (1.6)	185 (2.52)	*0.03*
	Non-Hispanic	7,088 (94.14)	179 (98.4)	6,909 (94.04)	
	Unknown or not reported	253 (3.36)	0 (0.0)	253 (3.44)	
Smoking status					
	Current smoker	3,893 (51.71)	112 (61.5)	3,781 (51.46)	*0.03*
	Former smoker	1,267 (16.83)	25 (13.7)	1,242 (16.91)	
	Never smoker or unknown	2,369 (31.47)	45 (24.7)	2,324 (31.63)	
Insurance					
	Private	4,642 (61.65)	110 (60.4)	4,532 (61.69)	0.79
	Public	6,901 (91.66)	179 (98.4)	6,722 (91.49)	*0.002*
	Self-paid or charity	540 (7.17)	41 (22.5)	499 (6.79)	*<0.001*
Number of years from the first encounter related to COPD in the data set					
	≤3	5,154 (68.46)	81 (44.5)	5,073 (69.05)	*<0.001*
	>3	2,375 (31.54)	101 (55.5)	2,274 (30.95)	
COPD medication prescription					
	Inhaled corticosteroid (ICS)	2,635 (35.00)	98 (53.8)	2,537 (34.53)	*<0.001*
	Short-acting muscarinic antagonist (SAMA)	1,202 (15.96)	68 (37.4)	1,134 (15.43)	*<0.001*
	Short-acting beta-2 agonist (SABA)	4,241 (56.33)	158 (86.8)	4,083 (55.57)	*<0.001*
	SABA and SAMA combination	1,809 (24.03)	115 (63.2)	1,694 (23.06)	*<0.001*
	Long-acting muscarinic antagonist (LAMA)	2,061 (27.37)	110 (60.4)	1,951 (26.56)	*<0.001*
	Long-acting beta-2 agonist (LABA)	1,760 (23.38)	77 (42.3)	1,683 (22.91)	*<0.001*
	LABA and LAMA combination	400 (5.31)	12 (6.6)	388 (5.28)	0.54
	ICS and LABA combination	1,804 (23.96)	75 (41.2)	1,729 (23.53)	*<0.001*
	ICS, LABA, and LAMA combination	69 (0.92)	1 (0.5)	68 (0.93)	0.90
	Phosphodiesterase-4 inhibitor	26 (0.35)	2 (1.1)	24 (0.33)	0.27
	Systemic corticosteroid	2,385 (31.68)	103 (56.6)	2,282 (31.06)	*<0.001*
Comorbidity					
	Allergic rhinitis	410 (5.45)	14 (7.7)	396 (5.39)	0.24
	Anxiety or depression	2,153 (28.60)	63 (34.6)	2,090 (28.45)	0.08
	Asthma	1,096 (14.56)	43 (23.6)	1,053 (14.33)	*<0.001*
	Congestive heart failure	1,412 (18.75)	43 (23.6)	1,369 (18.63)	0.11
	Diabetes	1,689 (22.43)	40 (22.0)	1,649 (22.44)	0.95
	Eczema	258 (3.43)	11 (6.0)	247 (3.36)	0.08
	Gastroesophageal reflux	1,443 (19.17)	47 (25.8)	1,396 (19.00)	*0.03*
	Hypertension	3,791 (50.35)	105 (57.7)	3,686 (50.17)	0.05
	Ischemic heart disease	1,658 (22.02)	54 (29.7)	1,604 (21.83)	*0.02*
	Lung cancer	203 (2.70)	3 (1.6)	200 (2.72)	0.51
	Obesity	669 (8.89)	21 (11.5)	648 (8.82)	0.25
	Sinusitis	279 (3.71)	7 (3.8)	272 (3.70)	0.99
	Sleep apnea	915 (12.15)	28 (15.4)	887 (12.07)	0.22

In the training set of the main analysis, most patient characteristics exhibited statistically significantly different distributions between the data instances linked to severe COPD exacerbations in the next year and those linked to no severe COPD exacerbation in the next year. Exceptions occurred on the patient characteristics of having prescriptions of inhaled corticosteroid (ICS), long-acting beta-2 agonist (LABA), and long-acting muscarinic antagonist (LAMA) combinations (*P*=.66); having prescriptions of phosphodiesterase-4 inhibitor (*P*=.06); presence of diabetes (*P*=.43), presence of eczema (*P*=.30); presence of lung cancer (*P*=.31); and presence of sinusitis (*P*=.61). In the test set of the main analysis, most patient characteristics exhibited statistically significantly different distributions between the data instances linked to severe COPD exacerbations in the next year and those linked to no severe COPD exacerbation in the next year. Exceptions occurred on the patient characteristics of having private insurance (*P*=.79); having prescriptions of LABA and LAMA combinations (*P*=.54); having prescriptions of

ICS, LABA, and LAMA combinations (*P*=.90); having prescriptions of phosphodiesterase-4 inhibitor (*P*=.27); presence of allergic rhinitis (*P*=.24); presence of anxiety or depression (*P*=.08); presence of congestive heart failure (*P*=.11); presence of diabetes (*P*=.95); presence of eczema (*P*=.08); presence of hypertension (*P*=.05); presence of lung cancer (*P*=.51); presence of obesity (*P*=.25); presence of sinusitis (*P*=.99); and presence of sleep apnea (*P*=.22).

Classification algorithm and features used in the final model

The XGBoost algorithm was chosen by our automatic machine learning model selection method [78]. As a tree-based algorithm, XGBoost handles missing values in the features naturally. As detailed in Hastie *et al*. [83], XGBoost automatically calculates an importance value for each feature based on the feature's apportioned contribution to the model. In the main analysis, the final model was created using XGBoost and the 229 features shown in descending order of their importance values in Table 2 of the Appendix. The other features that contributed no extra predictive power were automatically dropped by XGBoost.

Model performance in the main analysis

In the main analysis with the test set, the final model had an AUC of 0.866 (95% CI 0.838-0.892), as computed from the model's receiver operating characteristic curve (Figure 2). The model's performance measures varied with the cutoff threshold for binary classification (Table 6). When using the top 10.00% (752/7,529) of patients with the largest predicted risk to set the cutoff threshold for binary classification, the model had an accuracy of 90.33% (6,801/7,529, 95% CI 89.61-91.01), a sensitivity of 56.6% (103/182, 95% CI 49.2-64.2), a specificity of 91.17% (6,698/7,347, 95% CI 90.51-91.83), a PPV of 13.7% (103/752, 95% CI 11.2-16.2), and an NPV of 98.83% (6,698/6,777, 95% CI 98.55-99.08), as computed from the corresponding confusion matrix of the model (Table 7).

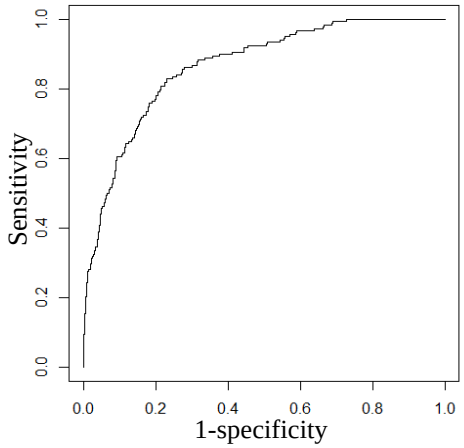

Figure 2. The receiver operating characteristic curve of the final model in the main analysis.

Table 6. In the main analysis, the performance measures of the final model with respect to using varying cutoff thresholds for binary classification.

Top percentage of patients with the largest predicted risk (%)	Accuracy (*N*=7,529), *n* (%)	Sensitivity (*N*=182), *n* (%)	Specificity (*N*=7,347), *n* (%)	Positive predictive value		Negative predictive value	
				n (%)	*N*	*n* (%)	*N*
1	7,336 (97.44)	32 (17.6)	7,304 (99.41)	32 (42.67)	75	7,304 (97.99)	7,454
2	7,299 (96.95)	51 (28.0)	7,248 (98.65)	51 (34.00)	150	7,248 (98.22)	7,379
3	7,236 (96.11)	57 (31.3)	7,179 (97.71)	57 (25.33)	225	7,179 (98.29)	7,304
4	7,170 (95.23)	62 (34.1)	7,108 (96.75)	62 (20.60)	301	7,108 (98.34)	7,228
5	7,111 (94.45)	70 (38.5)	7,041 (95.84)	70 (18.62)	376	7,041 (98.43)	7,153
6	7,062 (93.80)	83 (45.6)	6,979 (94.99)	83 (18.40)	451	6,979 (98.60)	7,078
7	6,994 (92.89)	87 (47.8)	6,907 (94.01)	87 (16.51)	527	6,907 (98.64)	7,002
8	6,927 (92.00)	91 (50.0)	6,836 (93.04)	91 (15.12)	602	6,836 (98.69)	6,927
9	6,860 (91.11)	95 (52.2)	6,765 (92.08)	95 (14.03)	677	6,765 (98.73)	6,852
10	6,801 (90.33)	103 (56.6)	6,698 (91.17)	103 (13.70)	752	6,698 (98.83)	6,777

15	6,458 (85.78)	120 (65.9)	6,338 (86.27)	120 (10.63)	1,129	6,338 (99.03)	6,400
20	6,118 (81.26)	138 (75.8)	5,980 (81.39)	138 (9.17)	1,505	5,980 (99.27)	6,024
25	5,767 (76.60)	151 (83.0)	5,616 (76.44)	151 (8.02)	1,882	5,616 (99.45)	5,647

Table 7. In the main analysis, the confusion matrix of the final model when using the top 10.00% (794/7,944) of patients with the largest predicted risk to set the cutoff threshold for binary classification.

Outcome class	Severe COPD exacerbations in the next year	No severe COPD exacerbation in the next year
Predicted severe COPD exacerbations in the next year	103	649
Predicted no severe COPD exacerbation in the next year	79	6,698

Recall that 27 candidate features were computed on ≥2 years of data. When we ignored these features and considered only those computed with the data in the index year, the model's AUC dropped from 0.866 to 0.859 (95% CI 0.834-0.884). The top 19 features shown in Table 2 of the Appendix have importance values ≥1%. When using only these features, the model's AUC dropped from 0.866 to 0.862 (95% CI 0.837-0.887). In this case, when using the top 10.00% (752/7,529) of patients with the largest predicted risk to set the cutoff threshold for binary classification, the model had an accuracy of 90.25% (6,795/7,529, 95% CI 89.56-90.90), a sensitivity of 54.9% (100/182, 95% CI 47.8-61.9), a specificity of 91.13% (6,695/7,347, 95% CI 90.43-91.78), a PPV of 13.3% (100/752, 95% CI 10.9-15.7), and an NPV of 98.79 (6,695/6,777, 95% CI 98.52-99.06).

Performance stability analysis

The final model in the main analysis and the model in the performance stability analysis had relatively similar performance (Table 8).

Table 8. The performance of the final model in the main analysis and the model in the performance stability analysis.

Performance measure	Final model in the main analysis		Model in the performance stability analysis	
	Value	95% CI	Value	95% CI
Accuracy	90.33% (6,801/7,529)	(89.61-91.01)	89.63% (6,354/7,089)	(88.94-90.32)
Sensitivity	56.6% (103/182)	(49.2-64.2)	46.3% (171/369)	(40.9-51.5)
Specificity	91.17% (6,698/7,347)	(90.51-91.83)	92.01% (6,183/6,720)	(91.36-92.69)
Positive predictive value	13.7% (103/752)	(11.2-16.2)	24.2% (171/708)	(20.8-27.2)
Negative predictive value	98.83% (6,698/6,777)	(98.55-99.08)	96.90% (6,183/6,381)	(96.43-97.31)
AUC	0.866	(0.838-0.892)	0.847	(0.828-0.864)

Discussion

Key findings

We created a machine learning model to predict severe COPD exacerbations in the next year in patients with COPD. The model had a higher AUC than the formerly published AUC of every prior model for predicting severe COPD exacerbations in the next year [20,25,27,28,30,33,35-43,46-49,51] (Table 9). After improving our model's performance measures further (e.g., by adding features extracted from clinical notes) and using our recently published automatic explanation method [84] to automatically explain the model's predictions, our model could be used as a decision support tool to advise care management's use for high-risk patients with COPD to improve outcomes.

Table 9. A comparison of our final model and several prior models to predict severe COPD exacerbations in patients with COPD. "—" means that the performance measure is unreported in the initial paper describing the model.

Model	Data	# of data instances	Prediction target (outcome)	Length of the period used to compute the outcome	Prevalence rate of the poor outcome	# of features checked	Classification algorithm	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	AUC
Our final model	Administrative and clinical	43,576	ED visit or inpatient stay for COPD	1 year	5.10%	278	XGBoost	56.6	91.17	13.7	98.83	0.866
Annavarapu *et al*. [20]	Administrative	45,722	Inpatient stay for COPD	1 year	11.63%	103	Logistic regression	17.3	97.5	48.1	90.0	0.77

Tavakoli *et al.* [21]	Administrative	222,219	Inpatient stay for COPD	2 months	1.02%	83	Gradient boosting	23	98	—	—	0.820
Samp *et al.* [22]	Administrative	478,772	Inpatient stay for COPD	6 months	2.2%	101	Logistic regression	17.6	96.6	—	—	—
Thomsen *et al.* [23]	Research	6,574	≥2 exacerbations (medication change or inpatient stay for COPD)	1 to 7 years	6.4%	11	Logistic regression	—	—	18	96	0.73
Orchard *et al.* [24]	Research	57,150	Inpatient stay for COPD	1 day	0.10%	153	Neural network	80	60	—	—	0.740
Suetomo *et al.* [25]	Research	123	Inpatient stay for COPD	1 year	12.2%	18	Logistic regression	53	49	—	—	0.79
Lee *et al.* [26]	Research and clinical	545	Medication change, ED visit, or inpatient stay for COPD	6 months	46%	10	Logistic regression	52	69	—	—	0.63
Faganello *et al.* [27]	Research	120	Outpatient, inpatient, or ED encounter for COPD	1 year	50%	16	Logistic regression	58.3	73.3	—	—	0.686
Alcázar *et al.* [28]	Research	127	Inpatient stay for COPD	1 year	39.4%	9	Logistic regression	76.2	77.3	61.5	87.2	0.809
Bertens *et al.* [29]	Research and clinical	1,033	Medication change or inpatient stay for COPD	2 years	28.3%	7	Logistic regression	—	—	—	—	0.66
Miravitlles *et al.* [30]	Research and clinical	713	Inpatient stay for COPD	1 year	22.2%	7	Logistic regression	—	—	—	—	0.582
Make *et al.* [31]	Research	3,141	Medication change, ED visit, or inpatient stay for COPD	6 months	—	38	Logistic regression	—	—	—	—	0.67
Montserrat-Capdevila *et al.* [32]	Administrative and clinical	2,501	Inpatient stay for COPD	3 years	32.5%	17	Logistic regression	—	—	—	—	0.72
Kerkhof *et al.* [33]	Research and clinical	16,565	≥2 exacerbations (medication change, ED visit, or inpatient stay for COPD)	1 year	19.6%	22	Logistic regression	—	—	—	—	0.735
Chen *et al.* [34]	Research	1,711	ED visit or inpatient stay for COPD	5 years	30.6%	14	Cox proportional hazard regression	—	—	—	—	0.74
Yii *et al.* [35]	Administrative and clinical	237	Inpatient stay for COPD	1 year	1.41 per patient-year	31	Negative binomial regression	—	—	—	—	0.789

Adibi *et al.* [36]	Research	2,380	ED visit or inpatient stay for COPD	1 year	0.29 per year	13	Mixed-effect logistic	—	—	—	—	0.77
Stanford *et al.* [37]	Administrative	258,668	Inpatient stay for COPD	1 year	8.5%	30	Logistic regression	—	—	—	—	0.749
Stanford *et al.* [38]	Administrative	223,824	Inpatient stay for COPD	1 year	6.63%	30	Logistic regression	—	—	—	—	0.711
Stanford *et al.* [39]	Administrative	92,496	Inpatient stay for COPD	1 year	—	30	Logistic regression	—	—	—	—	0.801
Stanford *et al.* [40]	Administrative	60,776	Inpatient stay for COPD	1 year	19.16%	8	Logistic regression	—	—	—	—	0.742
Jones *et al.* [41]	Clinical	375	Inpatient stay for COPD	1 year	—	4	Index	—	—	—	—	0.755
Jones *et al.* [42]	Research and clinical	7,105	Inpatient stay for COPD	1 year	—	8	Negative binomial regression	—	—	—	—	0.64
Fan *et al.* [43]	Research	3,282	Inpatient stay for COPD	1 year	4.3%	23	Logistic regression	—	—	—	—	0.706
Moy *et al.* [44]	Research and clinical	167	Inpatient stay for COPD	4 to 21 months	32.9%	6	Negative binomial regression	—	—	—	—	0.69
Briggs *et al.* [45]	Research	8,802	Inpatient stay for COPD	6 months to 3 years	9.0%	13	Cox proportional hazard regression	—	—	—	—	0.71
Lange *et al.* [46]	Administrative and research	6,628	Medication change or inpatient stay for COPD	1 year	4.8%	3	GOLD stratification	—	—	—	—	0.7
Abascal-Bolado *et al.* [47]	Research and clinical	493	Inpatient stay for COPD	1 year	—	8	Classification and regression tree	—	—	—	—	0.70
Blanco-Aparicio *et al.* [48]	Research	100	ED visit for COPD	1 year	21%	12	Logistic regression	—	—	—	—	0.651
Yoo *et al.* [49]	Research and clinical	260	Medication change, ED visit, or inpatient stay for COPD	1 year	40.8%	17	Logistic regression	—	—	—	—	0.69
Niewoehner *et al.* [50]	Research and clinical	1,829	Inpatient stay for COPD	6 months	8.3%	27	Cox proportional hazard regression	—	—	—	—	0.73
Austin *et al.* [51]	Administrative	638,926	COPD-related inpatient stay	1 year	—	34	Logistic regression	—	—	—	—	0.778
Marin *et al.* [52]	Research	275	Inpatient stay for COPD	6 months to 8 years	—	4	Logistic regression	86	73	—	—	0.88
Marin *et al.* [52]	Research	275	ED visit for COPD	6 months to 8 years	—	4	Logistic regression	58	87	—	—	0.78
Ställberg *et al.* [53]	Administrative and clinical	7,823	COPD-related inpatient stay	10 days	—	>4,000	XGBoost	16	—	11	—	0.86

In Table 2 of the Appendix, many of the top 19 features match the published (risk) factors that were highly correlated with COPD exacerbations, such as prior COPD exacerbations [18,60], prior healthcare encounters related to COPD [28,50], COPD medication usage [50], body mass index [70], peripheral capillary oxygen saturation [28], and heart rate [71].

We examined 278 candidate features, 82.4% (229/278) of which were used in the final model. Many omitted features are correlated with the outcome, but provided no extra predictive power on the UWM data set beyond the 229 features used in the final model.

The prevalence rate of severe COPD exacerbations had a sudden drop in 2019. Despite this drop, our model still showed reasonably robust performance over time. This is desired for clinical decision support.

Comparison with prior work

Researchers formerly created several models to predict severe COPD exacerbations in patients with COPD [20-53]. Table 9 presents comparisons between our final model and these models, which include all related models listed in Guerra *et al.*'s systematic reviews [85,86] as well as several recent models that were published after the reviews. Our final model predicted severe COPD exacerbations in the next year. Every prior model for predicting severe COPD exacerbations in the next year had an AUC that is ≤0.809, i.e., at least 0.057 lower than our final model's AUC. Compared with the prior models other than Ställberg *et al.*'s model [53] for predicting severe COPD exacerbations, our final model used more extensive features with predictive power, which helped improve model performance.

Our final model's prediction target covered both future ED visits and future inpatient stays for COPD, which we want to use care management to prevent. Among all prior models, only 2 [34,36] had prediction targets covering both future ED visits and future inpatient stays for COPD. Most of the prior models predicted either only future ED visits [48,52] or only future inpatient stays for COPD [20-22,24,25,28,30,32,35,37-45,47,50-52]. This would be insufficient for preventing both future ED visits and future inpatient stays for COPD. The other prior models [23,26,27,29,31,33,46,49] had prediction targets covering both moderate and severe COPD exacerbations, with moderate COPD exacerbations typically referring to COPD medication change such as the use of systemic corticosteroids. Those prediction targets were not specific enough for identifying patients at the highest risk for care management, as a care management program can host only a small portion of patients [17].

To make it suitable for use in daily clinical practice, our final model was built on routinely available administrative and clinical data. In comparison, Thomsen *et al.*'s models [23-31,33,34,36,42-50,52] used research data, some of which are unavailable in usual clinical practice. Thus, these models would be unsuitable for daily clinical use.

Our predictive model was developed to guide COPD care management's enrollment decisions and to prevent severe COPD exacerbations. To give enough lead time for preventive interventions to be effective and to use precious care management resources well, we chose severe COPD exacerbation in the next year as the prediction target. In comparison, Orchard *et al.*'s model [24] predicted inpatient stays for COPD in the next day. If a patient will incur an inpatient stay for COPD tomorrow, intervening starting from today could be too late to avoid the inpatient stay. At present, we are aware of no published conclusion on how long it will take for any intervention to be effective at preventing severe COPD exacerbations. In Longman *et al.*'s study [87,88], several clinicians had expressed the opinion that it could take as long as 3 months for any intervention to be effective at preventing inpatient stays for a chronic, ambulatory care sensitive condition. Our final model will have a different clinical use from the models that make short-term predictions. Foreseeing a severe COPD exacerbation in the next 12 months would be useful for identifying and personalizing medium-term interventions and maintenance therapies to change the course of the disease. In comparison, foreseeing a severe COPD exacerbation in the next 1 or few days can be useful for deciding acute management approaches to improve outcomes, such as preemptive hospitalization of the patient to avoid more severe adverse outcomes, but would be inadequate for trying to improve the course of the disease in a short amount of time. In fact, treatment approaches proven to be effective at reducing severe COPD exacerbations are usually not indicated for acute management.

Marin *et al.* [52] built a model to predict inpatient stays for COPD in up to the next 8 years with an AUC of 0.88, and a separate model to predict ED visits for COPD in up to the next 8 years with an AUC of 0.78. An inpatient stay or an ED visit that will happen several years later is too remote to be worth using precious care management resources now to prevent.

For the patients with COPD who will have severe COPD exacerbations in the future, sensitivity is the proportion of them whom the model identifies. The difference in sensitivity could greatly impact hospital use. Our final model's sensitivity is higher than the sensitivities achieved by Annavarapu *et al.*'s models [20-22,25,26,53]. Compared with our final model, Orchard *et al.*'s models [24,27,28] each reached a higher sensitivity at the price of a much lower specificity. For each of these 3 models, if we adjust the cutoff threshold for binary classification and make our final model have the same specificity as that model, our final model would achieve a higher sensitivity than that model. More specifically, at a specificity of 60.02% (4,410/7,347), our final model achieved a sensitivity of 90.1% (164/182), whereas Orchard *et al.*'s model [24] achieved a sensitivity of 80%. At a specificity of 73.30% (5,385/7,347), our final model achieved a sensitivity of 84.1% (153/182), whereas Faganello *et al.*'s model [27] achieved a sensitivity of 58.3%. At a specificity of 77.34% (5,682/7,347), our final model achieved a sensitivity of 81.9% (149/182), whereas Alcázar *et al.*'s model [28] achieved a sensitivity of 76.2%.

The prevalence rate of poor outcomes has a large impact on any model's PPV [89]. On our data set where this prevalence rate is around 5%, our final model reached a PPV of <14%. In comparison, on a data set where this prevalence rate is 11.63%, Annavarapu *et al.*'s model [20] reached a PPV of 48.1%. On a data set where this prevalence rate is 6.4%, Thomsen *et al.*'s model [23] reached a PPV of 18%. On a data set where this prevalence rate is 39.4%, Alcázar *et al.*'s model [28] reached a PPV of 61.5%. In all 3 cases, the higher prevalence rates of poor outcomes permitted the PPV to be larger.

Our data set is imbalanced, with only a small portion of patients to have severe COPD exacerbations in the next year. For imbalanced data sets, the area under the precision-recall curve (AUPRC) is a better measure of overall model performance than the AUC [90]. The AUPRC was reported for only Ställberg *et al.*'s model [53] among all of the prior models. Although Ställberg *et al.*'s model [53] had an AUC of 0.86 that is only slightly lower than our final model's AUC, our final model had an AUPRC of 0.24 (95% CI 0.18-0.31) that is 3 times as large as the 0.08 AUPRC of that model. In addition, that model predicted COPD-related inpatient stays, for which COPD can be any of the diagnoses, in the next 10 days. If a patient will incur an inpatient stay in the next 10 days, intervening starting from today could be too late to avoid the inpatient stay. In comparison, our final model predicted ED visits or inpatient stays with a principal diagnosis of COPD in the next year, allowing more lead time for preventive interventions to be effective.

Considerations for future clinical use

Our final model reached an AUC that is larger than every AUC formerly reported in the literature for predicting severe COPD exacerbations in the next year. Despite having a relatively low PPV, our final model could still benefit healthcare for 3 reasons.

First, healthcare systems like the UWM and Intermountain Healthcare use proprietary models, which have similar performance to the formerly published models, to allocate COPD care management resources. Our final model had a higher AUC than all formerly reported AUCs for predicting severe COPD exacerbations in the next year. Hence, although we plan to investigate using various techniques to further improve model performance in the future, we think it is already worth considering using our final model to replace the proprietary models currently being used at healthcare systems like the UWM for COPD care management.

Second, we set the cutoff threshold for binary classification at the top 10% (752/7,529) of patients with the largest predicted risk. In this case, a perfect model would achieve the theoretically maximum possible PPV of 24.2% (182/752). Our final model's PPV is 56.6% (103/182) of the theoretically maximum possible PPV. In other words, our final model captured 56.6% (103/182) of the patients with COPD who would have severe COPD exacerbations in the next year. If we change the cutoff threshold to the top 25% of patients with the largest predicted risk, the final model would capture 83.0% (151/182) of the patients with COPD who would have severe COPD exacerbations in the next year.

Third, a PPV at the level of our final model's PPV is suitable for identifying high-risk patients with COPD for low-cost preventive interventions, such as arranging a nurse to further follow-up with the patient via phone calls, teaching the patient to correctly use a COPD inhaler, teaching the patient the correct use of a peak flow meter to self-monitor symptoms at home, and enrolling the patient in a home-based pulmonary rehabilitation program [91].

Our final model used 229 features. To ease clinical deployment, we could reduce features, e.g., to the top 19 with importance values ≥1%. A feature's importance value differs across healthcare systems. If conditions permit, we should use a data set from the target healthcare system to compute the features' importance values and decide which features to retain.

Our final model was based on XGBoost [76], which leverages the hyper-parameter scale_pos_weight to balance the weights of the 2 outcome classes in our data set [92]. The scale_pos_weight hyper-parameter was set by our automatic model selection method [78] to a non-default value to maximize our final model's AUC [93]. This caused the side effect of greatly increasing our model's predicted probabilities of having future severe COPD exacerbations to values much larger than the true probabilities [92]. However, it does not affect our ability to identify the top portion of patients with the largest predicted risk for preventive interventions. If preferred, we could forgo the balancing by keeping scale_pos_weight at its default value 1. In this case, our model's AUC would drop by 0.003 to 0.863 (95% CI 0.835-0.888), which is still larger than every formerly published AUC for predicting severe COPD exacerbations in the next year.

Limitations

This study has several limitations that are worth future work.

First, this study used solely structured data. It is worth considering performing natural language processing to extract features from unstructured clinical notes to improve model performance. A model with higher performance can be used to better facilitate COPD care management.

Second, this study used age, diagnosis codes, and medication data to identify patients with COPD, and diagnosis codes and encounter information to define the prediction target. One can use age, diagnosis codes, and medication data to identify patients with COPD reasonably well [56]. Yet, diagnosis codes were shown to have a low sensitivity in capturing inpatient stays for COPD [94]. Our predictive model is likely to perform poorly at finding those patients who would experience only future

inpatient stays for COPD that are not captured by our current definition of the prediction target. We expect that this will not greatly affect our predictive model's usefulness for facilitating COPD care management. Based on our current definition of the prediction target, >5% of patients in our data set had severe COPD exacerbations in the following year. If fully captured by the predictive model, these patients would have already exceeded the service capacity of a typical care management program, which can take ≤3% of patients [17]. In the future, one could consider adding both medication data and information extracted from clinical notes via natural language processing to better capture inpatient stays for COPD.

Third, this study used non-deep learning classification algorithms. Deep learning has improved model performance for many clinical predictive modeling tasks [95-100]. It is worth investigating using deep learning to improve model performance for predicting severe COPD exacerbations.

Fourth, this study used data from one healthcare system: the UWM. It is worth evaluating our model's generalizability to other healthcare systems. We are working on obtaining a data set of patients with COPD from Intermountain Healthcare for this purpose [101].

Fifth, our data set contained no information on UWM patients' healthcare utilization at other healthcare systems. It is worth evaluating how our model's performance would change if data on UWM patients' healthcare use at other healthcare systems are available.

Conclusions

This work improved the state-of-the-art of predicting severe COPD exacerbations in patients with COPD. Particularly, our final model had a higher AUC than every formerly published model AUC on predicting severe COPD exacerbations in the next year. After improving our model's performance measures further and using our recently published automatic explanation method [84] to automatically explain the model's predictions, our model could be used in a decision support tool to guide care management's use for high-risk patients with COPD to improve outcomes.

Acknowledgments

GL and SZ were partially supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award Number R01HL142503. SZ was also partially supported by the National Library of Medicine Training Grant under Award Number T15LM007442. MA was partially supported by grants from the Flight Attendant Medical Research Institute (CIA190001) and the California Tobacco-Related Disease Research Program (T29IR0715). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. YT did the work at the University of Washington when she was a visiting PhD student.

Authors' contributions

GL and SZ were mainly responsible for the paper. SZ performed a literature review, extracted and analyzed the data, constructed the models, and wrote the first draft of the paper. GL conceptualized and designed the study, participated in doing data analysis, and rewrote the whole paper. MA and ZL provided clinical expertise, contributed to conceptualizing the presentation, and revised the paper. YT took part in extracting the data and identifying the biologically implausible values.

Conflicts of interest

None declared.

Abbreviations:

AUC: area under the receiver operating characteristic curve
AUPRC: area under the precision-recall curve
CI: confidence interval
COPD: chronic obstructive pulmonary disease
COVID-19: coronavirus disease 2019
ED: emergency department
FN: false negative
FP: false positive
ICD-9: International Classification of Diseases, Ninth Revision
ICD-10: International Classification of Diseases, Tenth Revision
ICS: inhaled corticosteroid
LABA: long-acting beta-2 agonist
LAMA: long-acting muscarinic antagonist
NPV: negative predictive value
PCP: primary care provider

PPV: positive predictive value
SABA: short-acting beta-2 agonist
SAMA: short-acting muscarinic antagonist
TN: true negative
TP: true positive
Weka: Waikato Environment for Knowledge Analysis
UWM: University of Washington Medicine
XGBoost: extreme gradient boosting

References

1. Ford ES, Murphy LB, Khavjou O, Giles WH, Holt JB, Croft JB. Total and state-specific medical and absenteeism costs of COPD among adults aged ≥ 18 years in the United States for 2010 and projections through 2020. Chest 2015 Jan;147(1):31-45. PMID:25058738
2. Centers for Disease Control and Prevention. Disease or Condition of the Week - COPD. 2019. https://www.cdc.gov/dotw/copd/index.html.
3. Global Initiative for Chronic Obstructive Lung Disease - GOLD. 2020 Gold Reports. 2020. https://goldcopd.org/gold-reports.
4. Blanchette CM, Dalal AA, Mapel D. Changes in COPD demographics and costs over 20 years. J Med Econ 2012;15(6):1176-1182. PMID:22812689
5. Anzueto A, Leimer I, Kesten S. Impact of frequency of COPD exacerbations on pulmonary function, health status and clinical outcomes. Int J Chron Obstruct Pulmon Dis 2009;4:245-251. PMID:19657398
6. Connors AF Jr, Dawson NV, Thomas C, Harrell FE Jr, Desbiens N, Fulkerson WJ, Kussin P, Bellamy P, Goldman L, Knaus WA. Outcomes following acute exacerbation of severe chronic obstructive lung disease. The SUPPORT investigators (Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments). Am J Respir Crit Care Med 1996 Oct;154(4 Pt 1):959-967. PMID:8887592
7. Viglio S, Iadarola P, Lupi A, Trisolini R, Tinelli C, Balbi B, Grassi V, Worlitzsch D, Döring G, Meloni F, Meyer KC, Dowson L, Hill SL, Stockley RA, Luisetti M. MEKC of desmosine and isodesmosine in urine of chronic destructive lung disease patients. Eur Respir J 2000 Jun;15(6):1039-1045. PMID:10885422
8. Kanner RE, Anthonisen NR, Connett JE; Lung Health Study Research Group. Lower respiratory illnesses promote FEV(1) decline in current smokers but not ex-smokers with mild chronic obstructive pulmonary disease: results from the lung health study. Am J Respir Crit Care Med 2001 Aug 1;164(3):358-364. PMID:11500333
9. Spencer S, Jones PW; GLOBE Study Group. Time course of recovery of health status following an infective exacerbation of chronic bronchitis. Thorax 2003 Jul;58(7):589-593. PMID:12832673
10. Spencer S, Calverley PM, Sherwood Burge P, Jones PW; ISOLDE Study Group. Inhaled steroids in obstructive lung disease. Health status deterioration in patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2001 Jan;163(1):122-128. PMID:11208636
11. Johnston J, Longman J, Ewald D, King J, Das S, Passey M. Study of potentially preventable hospitalisations (PPH) for chronic conditions: what proportion are preventable and what factors are associated with preventable PPH? BMJ Open 2020;10(11):e038415. PMID:33168551
12. Billings J, Zeitel L, Lukomnik J, Carey TS, Blank AE, Newman L. Impact of socioeconomic status on hospital use in New York City. Health Aff (Millwood) 1993 Spring;12(1):162-173. PMID:8509018
13. Mays GP, Claxton G, White J. Managed care rebound? Recent changes in health plans' cost containment strategies. Health Aff (Millwood) 2004;Suppl Web Exclusives:W4-427-436. PMID:15451964
14. Rice KL, Dewan N, Bloomfield HE, Grill J, Schult TM, Nelson DB, Kumari S, Thomas M, Geist LJ, Beaner C, Caldwell M, Niewoehner DE. Disease management program for chronic obstructive pulmonary disease: a randomized controlled trial. Am J Respir Crit Care Med 2010 Oct 1;182(7):890-896. PMID:20075385
15. Bandurska E, Damps-Konstańska I, Popowski P, Jędrzejczyk T, Janowiak P, Świętnicka K, Zarzeczna-Baran M, Jassem E. Impact of integrated care model (ICM) on direct medical costs in management of advanced chronic obstructive pulmonary disease (COPD). Med Sci Monit 2017 Jun 12;23:2850-2862. PMID:28603270
16. Curry N, Billings J, Darin B, Dixon J, Williams M, Wennberg D. Predictive Risk Project Literature Review. London: King's Fund. http://www.kingsfund.org.uk/sites/files/kf/field/field_document/predictive-risk-literature-review-june2005.pdf, 2005.
17. Axelrod RC, Vogel D. Predictive modeling in health plans. Dis Manag Health Outcomes 2003;11(12):779-787. doi:10.2165/00115677-200311120-00003
18. Hurst JR, Vestbo J, Anzueto A, Locantore N, Müllerova H, Tal-Singer R, Miller B, Lomas DA, Agusti A, Macnee W, Calverley P, Rennard S, Wouters EF, Wedzicha JA; Evaluation of COPD Longitudinally to Identify Predictive Surrogate

Endpoints (ECLIPSE) Investigators. Susceptibility to exacerbation in chronic obstructive pulmonary disease. N Engl J Med 2010 Sep 16;363(12):1128-1138. PMID:20843247
19. Blagev DP, Collingridge DS, Rea S, Press VG, Churpek MM, Carey K, Mularski RA, Zeng S, Arjomandi M. Stability of frequency of severe chronic obstructive pulmonary disease exacerbations and health care utilization in clinical populations. Chronic Obstr Pulm Dis 2018 Jun 20;5(3):208-220. PMID:30584584
20. Annavarapu S, Goldfarb S, Gelb M, Moretz C, Renda A, Kaila S. Development and validation of a predictive model to identify patients at risk of severe COPD exacerbations using administrative claims data. Int J Chron Obstruct Pulmon Dis 2018 Jul 11;13:2121-2130. PMID:30022818
21. Tavakoli H, Chen W, Sin DD, FitzGerald JM, Sadatsafavi M. Predicting severe chronic obstructive pulmonary disease exacerbations. Developing a population surveillance approach with administrative data. Ann Am Thorac Soc 2020 Sep;17(9):1069-1076. PMID:32383971
22. Samp JC, Joo MJ, Schumock GT, Calip GS, Pickard AS, Lee TA. Predicting acute exacerbations in chronic obstructive pulmonary disease. J Manag Care Spec Pharm 2018 Mar;24(3):265-279. PMID:29485951
23. Thomsen M, Ingebrigtsen TS, Marott JL, Dahl M, Lange P, Vestbo J, Nordestgaard BG. Inflammatory biomarkers and exacerbations in chronic obstructive pulmonary disease. JAMA 2013 Jun 12;309(22):2353-2361. PMID:23757083
24. Orchard P, Agakova A, Pinnock H, Burton CD, Sarran C, Agakov F, McKinstry B. Improving prediction of risk of hospital admission in chronic obstructive pulmonary disease: application of machine learning to telemonitoring data. J Med Internet Res 2018 Sep 21;20(9):e263. PMID:30249589
25. Suetomo M, Kawayama T, Kinoshita T, Takenaka S, Matsuoka M, Matsunaga K, Hoshino T. COPD assessment tests scores are associated with exacerbated chronic obstructive pulmonary disease in Japanese patients. Respir Investig 2014 Sep;52(5):288-295. PMID:25169844
26. Lee SD, Huang MS, Kang J, Lin CH, Park MJ, Oh YM, Kwon N, Jones PW, Sajkov D; Investigators of the Predictive Ability of CAT in Acute Exacerbations of COPD (PACE) Study. The COPD assessment test (CAT) assists prediction of COPD exacerbations in high-risk patients. Respir Med 2014 Apr;108(4):600-608. PMID:24456695
27. Faganello MM, Tanni SE, Sanchez FF, Pelegrino NR, Lucheta PA, Godoy I. BODE index and GOLD staging as predictors of 1-year exacerbation risk in chronic obstructive pulmonary disease. Am J Med Sci 2010 Jan;339(1):10-14. PMID:19926966
28. Alcázar B, García-Polo C, Herrejón A, Ruiz LA, de Miguel J, Ros JA, García-Sidro P, Conde GT, López-Campos JL, Martínez C, Costán J, Bonnin M, Mayoralas S, Miravitlles M. Factors associated with hospital admission for exacerbation of chronic obstructive pulmonary disease. Arch Bronconeumol 2012 Mar;48(3):70-76. PMID:22196478
29. Bertens LC, Reitsma JB, Moons KG, van Mourik Y, Lammers JW, Broekhuizen BD, Hoes AW, Rutten FH. Development and validation of a model to predict the risk of exacerbations in chronic obstructive pulmonary disease. Int J Chron Obstruct Pulmon Dis 2013;8:493-499. PMID:24143086
30. Miravitlles M, Guerrero T, Mayordomo C, Sánchez-Agudo L, Nicolau F, Segú JL. Factors associated with increased risk of exacerbation and hospital admission in a cohort of ambulatory COPD patients: a multiple logistic regression analysis. The EOLO Study Group. Respiration 2000;67(5):495-501. PMID:11070451
31. Make BJ, Eriksson G, Calverley PM, Jenkins CR, Postma DS, Peterson S, Östlund O, Anzueto A. A score to predict short-term risk of COPD exacerbations (SCOPEX). Int J Chron Obstruct Pulmon Dis 2015 Jan 27;10:201-209. PMID:25670896
32. Montserrat-Capdevila J, Godoy P, Marsal JR, Barbé F. Predictive model of hospital admission for COPD exacerbation. Respir Care 2015 Sep;60(9):1288-1294. PMID:26286737
33. Kerkhof M, Freeman D, Jones R, Chisholm A, Price DB; Respiratory Effectiveness Group. Predicting frequent COPD exacerbations using primary care data. Int J Chron Obstruct Pulmon Dis 2015 Nov 9;10:2439-2450. PMID:26609229
34. Chen X, Wang Q, Hu Y, Zhang L, Xiong W, Xu Y, Yu J, Wang Y. A nomogram for predicting severe exacerbations in stable COPD patients. Int J Chron Obstruct Pulmon Dis 2020 Feb 18;15:379-388. PMID:32110006
35. Yii ACA, Loh CH, Tiew PY, Xu H, Taha AAM, Koh J, Tan J, Lapperre TS, Anzueto A, Tee AKH. A clinical prediction model for hospitalized COPD exacerbations based on "treatable traits". Int J Chron Obstruct Pulmon Dis 2019 Mar 27;14:719-728. PMID:30988606
36. Adibi A, Sin DD, Safari A, Johnson KM, Aaron SD, FitzGerald JM, Sadatsafavi M. The Acute COPD Exacerbation Prediction Tool (ACCEPT): a modelling study. Lancet Respir Med 2020 Oct;8(10):1013-1021. PMID:32178776
37. Stanford RH, Nag A, Mapel DW, Lee TA, Rosiello R, Vekeman F, Gauthier-Loiselle M, Duh MS, Merrigan JF, Schatz M. Validation of a new risk measure for chronic obstructive pulmonary disease exacerbation using health insurance claims data. Ann Am Thorac Soc 2016 Jul;13(7):1067-1075. PMID:27070274
38. Stanford RH, Nag A, Mapel DW, Lee TA, Rosiello R, Schatz M, Vekeman F, Gauthier-Loiselle M, Merrigan JFP, Duh MS. Claims-based risk model for first severe COPD exacerbation. Am J Manag Care 2018 Feb 1;24(2):e45-e53. PMID:29461849

39. Stanford RH, Lau MS, Li Y, Stemkowski S. External validation of a COPD risk measure in a commercial and Medicare population: the COPD treatment ratio. J Manag Care Spec Pharm 2019 Jan;25(1):58-69. PMID:30589629
40. Stanford RH, Korrer S, Brekke L, Reinsch T, Bengtson LGS. Validation and assessment of the COPD treatment ratio as a predictor of severe exacerbations. Chronic Obstr Pulm Dis 2020 Jan;7(1):38-48. PMID:31999901
41. Jones RC, Donaldson GC, Chavannes NH, Kida K, Dickson-Spillmann M, Harding S, Wedzicha JA, Price D, Hyland ME. Derivation and validation of a composite index of severity in chronic obstructive pulmonary disease: the DOSE Index. Am J Respir Crit Care Med 2009 Dec 15;180(12):1189-1195. PMID:19797160
42. Jones RC, Price D, Chavannes NH, Lee AJ, Hyland ME, Ställberg B, Lisspers K, Sundh J, van der Molen T, Tsiligianni I; UNLOCK Group of the IPCRG. Multi-component assessment of chronic obstructive pulmonary disease: an evaluation of the ADO and DOSE indices and the global obstructive lung disease categories in international primary care data sets. NPJ Prim Care Respir Med 2016 Apr 7;26:16010. PMID:27053297
43. Fan VS, Curtis JR, Tu SP, McDonell MB, Fihn SD; Ambulatory Care Quality Improvement Project Investigators. Using quality of life to predict hospitalization and mortality in patients with obstructive lung diseases. Chest 2002 Aug;122(2):429-436. PMID:12171813
44. Moy ML, Teylan M, Danilack VA, Gagnon DR, Garshick E. An index of daily step count and systemic inflammation predicts clinical outcomes in chronic obstructive pulmonary disease. Ann Am Thorac Soc 2014 Feb;11(2):149-157. PMID:24308588
45. Briggs A, Spencer M, Wang H, Mannino D, Sin DD. Development and validation of a prognostic index for health outcomes in chronic obstructive pulmonary disease. Arch Intern Med 2008 Jan 14;168(1):71-79. PMID:18195198
46. Lange P, Marott JL, Vestbo J, Olsen KR, Ingebrigtsen TS, Dahl M, Nordestgaard BG. Prediction of the clinical course of chronic obstructive pulmonary disease, using the new GOLD classification: a study of the general population. Am J Respir Crit Care Med 2012 Nov 15;186(10):975-981. PMID:22997207
47. Abascal-Bolado B, Novotny PJ, Sloan JA, Karpman C, Dulohery MM, Benzo RP. Forecasting COPD hospitalization in the clinic: optimizing the chronic respiratory questionnaire. Int J Chron Obstruct Pulmon Dis 2015 Oct 22;10:2295-2301. PMID:26543362
48. Blanco-Aparicio M, Vázquez I, Pita-Fernández S, Pértega-Diaz S, Verea-Hernando H. Utility of brief questionnaires of health-related quality of life (Airways Questionnaire 20 and Clinical COPD Questionnaire) to predict exacerbations in patients with asthma and COPD. Health Qual Life Outcomes 2013 May 27;11:85. PMID:23706146
49. Yoo JW, Hong Y, Seo JB, Chae EJ, Ra SW, Lee JH, Kim EK, Baek S, Kim TH, Kim WJ, Lee JH, Lee SM, Lee S, Lim SY, Shin TR, Yoon HI, Sheen SS, Lee JS, Huh JW, Oh YM, Lee SD. Comparison of clinico-physiologic and CT imaging risk factors for COPD exacerbation. J Korean Med Sci 2011 Dec;26(12):1606-1612. PMID:22147998
50. Niewoehner DE, Lokhnygina Y, Rice K, Kuschner WG, Sharafkhaneh A, Sarosi GA, Krumpe P, Pieper K, Kesten S. Risk indexes for exacerbations and hospitalizations due to COPD. Chest 2007 Jan;131(1):20-28. PMID:17218552
51. Austin PC, Stanbrook MB, Anderson GM, Newman A, Gershon AS. Comparative ability of comorbidity classification methods for administrative data to predict outcomes in patients with chronic obstructive pulmonary disease. Ann Epidemiol 2012 Dec;22(12):881-887. PMID:23121992
52. Marin JM, Carrizo SJ, Casanova C, Martinez-Camblor P, Soriano JB, Agusti AG, Celli BR. Prediction of risk of COPD exacerbations by the BODE index. Respir Med 2009 Mar;103(3):373-378. PMID:19013781
53. Ställberg B, Lisspers K, Larsson K, Janson C, Müller M, Łuczko M, Kjøller Bjerregaard B, Bacher G, Holzhauer B, Goyal P, Johansson G. Predicting hospitalization due to COPD exacerbations in Swedish primary care patients using machine learning - based on the ARCTIC study. Int J Chron Obstruct Pulmon Dis 2021 Mar 16;16:677-688. PMID:33758504
54. Tong Y, Liao ZC, Tarczy-Hornoch P, Luo G. Using a constraint-based method to identify chronic disease patients who are apt to obtain care mostly within a given health care system: retrospective cohort study. JMIR Form Res 2021;5(10):e26314. PMID:34617906
55. National Quality Forum. NQF #1891 Hospital 30-day, all-cause, risk-standardized readmission rate (RSRR) following chronic obstructive pulmonary disease (COPD) hospitalization. 2012. http://www.qualityforum.org/Projects/n-r/Pulmonary_Endorsement_Maintenance/1891_30_Day_RSRR_COPD.aspx.
56. Cooke CR, Joo MJ, Anderson SM, Lee TA, Udris EM, Johnson E, Au DH. The validity of using ICD-9 codes and pharmacy records to identify patients with chronic obstructive pulmonary disease. BMC Health Serv Res 2011 Feb 16;11:37. PMID:21324188
57. Nguyen HQ, Chu L, Amy Liu IL, Lee JS, Suh D, Korotzer B, Yuen G, Desai S, Coleman KJ, Xiang AH, Gould MK. Associations between physical activity and 30-day readmission risk in chronic obstructive pulmonary disease. Ann Am Thorac Soc 2014 Jun;11(5):695-705. PMID:24713094
58. Lindenauer PK, Grosso LM, Wang C, Wang Y, Krishnan JA, Lee TA, Au DH, Mularski RA, Bernheim SM, Drye EE. Development, validation, and results of a risk-standardized measure of hospital 30-day mortality for patients with exacerbation of chronic obstructive pulmonary disease. J Hosp Med 2013 Aug;8(8):428-435. PMID:23913593

59. Qureshi H, Sharafkhaneh A, Hanania NA. Chronic obstructive pulmonary disease exacerbations: latest evidence and clinical implications. Ther Adv Chronic Dis 2014 Sep;5(5):212-227. PMID:25177479
60. Müllerova H, Maselli DJ, Locantore N, Vestbo J, Hurst JR, Wedzicha JA, Bakke P, Agusti A, Anzueto A. Hospitalized exacerbations of COPD: risk factors and outcomes in the ECLIPSE cohort. Chest 2015 Apr;147(4):999-1007. PMID:25356881
61. Donaldson GC, Seemungal TA, Bhowmik A, Wedzicha JA. Relationship between exacerbation frequency and lung function decline in chronic obstructive pulmonary disease. Thorax 2002 Oct;57(10):847-852. PMID:12324669
62. Hurst JR, Donaldson GC, Quint JK, Goldring JJ, Baghai-Ravary R, Wedzicha JA. Temporal clustering of exacerbations in chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2009 Mar 1;179(5):369-374. PMID:19074596
63. Similowski T, Agustí A, MacNee W, Schönhofer B. The potential impact of anaemia of chronic disease in COPD. Eur Respir J 2006 Feb;27(2):390-396. PMID:16452598
64. Dahl M, Vestbo J, Lange P, Bojesen SE, Tybjaerg-Hansen A, Nordestgaard BG. C-reactive protein as a predictor of prognosis in chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2007 Feb 1;175(3):250-255. PMID:17053205
65. Hoenderdos K, Condliffe A. The neutrophil in chronic obstructive pulmonary disease. Am J Respir Cell Mol Biol 2013 May;48(5):531-539. PMID:23328639
66. Lonergan M, Dicker AJ, Crichton ML, Keir HR, Van Dyke MK, Mullerova H, Miller BE, Tal-Singer R, Chalmers JD. Blood neutrophil counts are associated with exacerbation frequency and mortality in COPD. Respir Res 2020 Jul 1;21(1):166. PMID:32611352
67. Chambellan A, Chailleux E, Similowski T; ANTADIR Observatory Group. Prognostic value of the hematocrit in patients with severe COPD receiving long-term oxygen therapy. Chest 2005 Sep;128(3):1201-1208. PMID:16162707
68. Toft-Petersen AP, Torp-Pedersen C, Weinreich UM, Rasmussen BS. Association between hemoglobin and prognosis in patients admitted to hospital for COPD. Int J Chron Obstruct Pulmon Dis 2016 Nov 10;11:2813-2820. PMID:27877035
69. van Dijk EJ, Vermeer SE, de Groot JC, van de Minkelis J, Prins ND, Oudkerk M, Hofman A, Koudstaal PJ, Breteler MM. Arterial oxygen saturation, COPD, and cerebral small vessel disease. J Neurol Neurosurg Psychiatry 2004 May;75(5):733-736. PMID:15090569
70. Kessler R, Faller M, Fourgaut G, Mennecier B, Weitzenblum E. Predictive factors of hospitalization for acute exacerbation in a series of 64 patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med 1999 Jan;159(1):158-164. PMID:9872834
71. Fermont JM, Masconi KL, Jensen MT, Ferrari R, Di Lorenzo VAP, Marott JM, Schuetz P, Watz H, Waschki B, Müllerova H, Polkey MI, Wilkinson IB, Wood AM. Biomarkers and clinical outcomes in COPD: a systematic review and meta-analysis. Thorax 2019 May;74(5):439-446. PMID:30617161
72. Halpin DM, Miravitlles M, Metzdorf N, Celli B. Impact and prevention of severe exacerbations of COPD: a review of the evidence. Int J Chron Obstruct Pulmon Dis 2017 Oct 5;12:2891-2908. PMID:29062228
73. Luo G, He S, Stone BL, Nkoy FL, Johnson MD. Developing a model to predict hospital encounters for asthma in asthmatic patients: secondary analysis. JMIR Med Inform 2020;8(1):e16080. PMID:31961332
74. Tong Y, Messinger AI, Wilcox AB, Mooney SD, Davidson GH, Suri P, Luo G. Forecasting future asthma hospital encounters of patients with asthma in an academic health care system: predictive model development and secondary analysis study. J Med Internet Res 2021;23(4):e22796. PMID:33861206
75. Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Burlington, MA: Morgan Kaufmann; 2016. ISBN:0128042915
76. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: KDD'16; August 13-17, 2016; San Francisco, CA p. 785-794. doi:10.1145/2939672.2939785
77. XGBoost JVM package. 2021. https://xgboost.readthedocs.io/en/latest/jvm/index.html.
78. Zeng X, Luo G. Progressive sampling-based Bayesian optimization for efficient and automatic machine learning model selection. Health Inf Sci Syst 2017;5(1):2. PMID:29038732
79. Thornton C, Hutter F, Hoos HH, Leyton-Brown K. Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013 Presented at: KDD'13; August 11-14, 2013; Chicago, IL p. 847-855. doi:10.1145/2487575.2487629
80. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, 2nd ed. New York, NY: Springer; 2019. ISBN:3030163989
81. Sykes DL, Faruqi S, Holdsworth L, Crooks MG. Impact of COVID-19 on COPD and asthma admissions, and the pandemic from a patient's perspective. ERJ Open Res 2021 Feb 8;7(1):00822-2020. PMID:33575313
82. Agresti A. Categorical Data Analysis, 3rd ed. Hoboken, NJ: Wiley; 2012. ISBN:9780470463635

83. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY: Springer; 2016. ISBN:0387848576
84. Luo G, Johnson MD, Nkoy FL, He S, Stone BL. Automatically explaining machine learning prediction results on asthma hospital visits in patients with asthma: secondary analysis. JMIR Med Inform 2020 Dec 31;8(12):e21965. PMID:33382379
85. Guerra B, Gaveikaite V, Bianchi C, Puhan MA. Prediction models for exacerbations in patients with COPD. Eur Respir Rev 2017 Jan 17;26(143):160061. PMID:28096287
86. Bellou V, Belbasis L, Konstantinidis AK, Tzoulaki I, Evangelou E. Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal. BMJ 2019 Oct 4;367:l5358. PMID:31585960
87. Longman JM, Passey ME, Ewald DP, Rix E, Morgan GG. Admissions for chronic ambulatory care sensitive conditions - a useful measure of potentially preventable admission? BMC Health Serv Res 2015 Oct 16;15:472. PMID:26475293
88. Johnston JJ, Longman JM, Ewald DP, Rolfe MI, Diez Alvarez S, Gilliland AHB, Chung SC, Das SK, King JM, Passey ME. Validity of a tool designed to assess the preventability of potentially preventable hospitalizations for chronic conditions. Fam Pract 2020 Jul 23;37(3):390-394. PMID:31848589
89. Ranganathan P, Aggarwal R. Common pitfalls in statistical analysis: Understanding the properties of diagnostic tests - Part 1. Perspect Clin Res 2018;9(1):40-43. PMID:29430417
90. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning. 2006 Presented at: ICML'06; June 25-29, 2006; Pittsburgh, PA p. 233-240. doi:10.1145/1143844.1143874
91. Burge AT, Holland AE, McDonald CF, Abramson MJ, Hill CJ, Lee AL, Cox NS, Moore R, Nicolson C, O'Halloran P, Lahham A, Gillies R, Mahal A. Home-based pulmonary rehabilitation for COPD using minimal resources: an economic analysis. Respirology 2020 Feb;25(2):183-190. PMID:31418515
92. XGBoost parameters. 2021. https://xgboost.readthedocs.io/en/latest/parameter.html.
93. Notes on parameter tuning. 2021. https://xgboost.readthedocs.io/en/latest/tutorials/param_tuning.html.
94. Stein BD, Bautista A, Schumock GT, Lee TA, Charbeneau JT, Lauderdale DS, Naureckas ET, Meltzer DO, Krishnan JA. The validity of International Classification of Diseases, Ninth Revision, Clinical Modification diagnosis codes for identifying patients hospitalized for COPD exacerbations. Chest 2012;141(1):87-93. PMID:21757568
95. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M, Sundberg P, Yee H, Zhang K, Zhang Y, Flores G, Duggan GE, Irvine J, Le Q, Litsch K, Mossin A, Tansuwan J, Wang D, Wexler J, Wilson J, Ludwig D, Volchenboum SL, Chou K, Pearson M, Madabushi S, Shah NH, Butte AJ, Howell M, Cui C, Corrado GS, Dean J. Scalable and accurate deep learning with electronic health records. npj Digital Medicine 2018;1:18. doi:10.1038/s41746-018-0029-1
96. Lipton ZC, Kale DC, Elkan C, Wetzel RC. Learning to diagnose with LSTM recurrent neural networks. In: Proceedings of the International Conference on Learning Representations. 2016 Presented at: International Conference on Learning Representations; May 2-4, 2016; San Juan, Puerto Rico p. 1-18. https://arxiv.org/abs/1511.03677
97. Kam HJ, Kim HY. Learning representations for the early detection of sepsis with deep neural networks. Comput Biol Med 2017;89:248-255. PMID:28843829
98. Razavian N, Marcus J, Sontag D. Multi-task prediction of disease onsets from longitudinal laboratory tests. In: Proceedings of the Machine Learning in Health Care Conference. 2016 Presented at: Machine Learning in Health Care Conference; August 19-20, 2016; Los Angeles, CA p. 73-100. http://proceedings.mlr.press/v56/Razavian16.html
99. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow PM, Zietz M, Hoffman MM, Xie W, Rosen GL, Lengerich BJ, Israeli J, Lanchantin J, Woloszynek S, Carpenter AE, Shrikumar A, Xu J, Cofer EM, Lavender CA, Turaga SC, Alexandari AM, Lu Z, Harris DJ, DeCaprio D, Qi Y, Kundaje A, Peng Y, Wiley LK, Segler MHS, Boca SM, Swamidass SJ, Huang A, Gitter A, Greene CS. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 2018;15(141):20170387. PMID:29618526
100. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for Electronic Health Record (EHR) analysis. IEEE J Biomed Health Inform 2018;22(5):1589-1604. PMID:29989977
101. Luo G, Stone BL, Koebnick C, He S, Au DH, Sheng X, Murtaugh MA, Sward KA, Schatz M, Zeiger RS, Davidson GH, Nkoy FL. Using temporal features to provide data-driven clinical early warnings for chronic obstructive pulmonary disease and asthma care management: protocol for a secondary analysis. JMIR Res Protoc 2019;8(6):e13783. PMID:31199308

Appendix

Table 1. The candidate features.

Feature category	Features	Notes
Patient demographics	Age [30,50,59]; gender; marital status (married, single, partnered, divorced, widowed, or separated); race; ethnicity (Hispanic or non-Hispanic); and language.	Acceptable range: Age: 40-122 years [102]. Number of features: 6.
Laboratory test	Minimum Alpha-1 antitrypsin (A1AT) level; maximum A1AT level; whether the minimum A1AT level is abnormally low; minimum arterial oxygen saturation (SaO_2) level [69]; maximum arterial partial pressure of carbon dioxide ($PaCO_2$) level; minimum $PaCO_2$ level; maximum arterial partial pressure of oxygen (PaO_2) level; minimum PaO_2 level; maximum blood eosinophil count; maximum percentage of blood eosinophils; maximum blood neutrophil count [65,66]; maximum percentage of blood neutrophils; maximum C-reactive protein (CRP) level [64]; whether the maximum CRP level is abnormally high; maximum hematocrit (Hct) level; minimum Hct level [67]; whether the maximum Hct level is abnormally high; whether the minimum Hct level is abnormally low; maximum hemoglobin A1c (HbA1c) level; maximum hemoglobin (Hgb) level [68]; minimum Hgb level; whether the maximum Hgb level is abnormally high; whether the minimum Hgb level is abnormally low; whether an immunoglobulin E (IgE) test was performed; maximum total serum IgE level; whether the maximum total serum IgE level is abnormally high; maximum red blood cell count [63]; maximum white blood cell count [18,60]; no. of laboratory tests; no. of days from the last laboratory test; and no. of laboratory tests having abnormal results.	Number of features: 31.
Vital sign	Maximum body mass index (BMI) [70]; the relative change of BMI = (the last logged BMI / the first logged BMI - 1) × 100%; maximum diastolic blood pressure; average diastolic blood pressure; maximum heart rate [71]; average heart rate; maximum height; minimum peak expiratory flow [61]; average peak expiratory flow; minimum peripheral capillary oxygen saturation (SpO_2) [28]; average SpO_2; maximum respiratory rate; average respiratory rate; maximum systolic blood pressure; average systolic blood pressure; maximum temperature; average temperature; and the relative change of weight = (the last logged weight / the first logged weight - 1) × 100%.	Acceptable ranges: Weight: 0.26-635 kilograms [103,104]; height: 0.24-2.72 meters [105,106]; BMI: 7.5-204.0 [107,108]; systolic and diastolic blood pressure: 0-300 mm Hg; heart rate: 30-300 beats/minute; respiratory rate: 0-120; temperature: 80-110 °F; SpO_2: 0-100; and peak expiratory flow: 0-700 liters/minute. Number of features: 18.
Spirometry	Whether a spirometry was performed; average forced expiratory volume in 1 second (FEV_1) [61]; minimum FEV_1; average FEV_1 percent predicted; minimum FEV_1 percent predicted; average forced vital capacity (FVC); minimum FVC; average FEV_1/FVC ratio; minimum FEV_1/FVC ratio; whether any FEV_1 percent predicted is <80% with a normal FEV_1/FVC ratio (preserved ratio impaired spirometry, PRISm); and the highest GOLD stage of COPD [3,18,30,50,60].	FEV_1 percent predicted was computed using the Hankinson's prediction equation [109-111]. Number of features: 11.
Diagnosis	Computed based on International Classification of Diseases, Ninth Revision (ICD-9) and International Classification of Diseases, Tenth Revision (ICD-10) diagnosis codes: No. of ICD-9 and ICD-10 diagnosis codes; no. of chronic obstructive pulmonary disease (COPD) diagnoses; no. of primary or principal COPD diagnoses; no. of diagnoses of acute COPD exacerbation; no. of years from the first encounter related to COPD [50]; whether the last COPD diagnosis is a primary or principal diagnosis; no. of days from the last COPD diagnosis; no. of days from the last diagnosis of acute COPD exacerbation; no. of diagnoses of	Number of features: 62.

	noncompliance with medication regimen; acquired immunodeficiency syndrome; allergic rhinitis; Alzheimer's or Parkinson's disease; anaphylaxis; anxiety or depression [72]; asthma; breathing abnormality like dyspnea; bronchopulmonary dysplasia; cirrhosis; cerebrovascular disease; congestive heart failure [30]; cystic fibrosis; decreased tone; dementia; diabetes without chronic complication [30]; diabetes with chronic complication; eczema; esophagitis; folate deficiency; gastroesophageal reflux [18]; gastrointestinal bleeding; gastrointestinal obstruction; gastrostomy tube; hypertension; immunoglobulin A (IgA) deficiency; increased tone; inflammatory bowel disease; ischemic heart disease [30]; lung cancer; malignancy; mental disorder; metastatic solid tumor; mild liver disease; moderate or severe liver disease; myocardial infarction; obesity; paraplegia or hemiplegia; peptic ulcer disease; peripheral vascular disease [50]; pneumonia; pregnancy; psoriasis; renal disease [30]; rheumatic disease; sleep apnea; substance use; upper respiratory tract infection; vasculitis; vitamin D deficiency; and vocal cord dysfunction. Computed based on ICD-9 and ICD-10 procedure and diagnosis codes: Tracheostomy. Computed based on Current Procedural Terminology (CPT) and Healthcare Common Procedure Coding System (HCPCS) procedure codes: Cataract; and sinusitis.	
Problem list	No. of active problems; no. of active problems of COPD; no. of active problems of COPD exacerbation; no. of active problems of anxiety or depression; no. of active problems of asthma; no. of active problems of asthma with (acute) exacerbations; no. of active problems of gastroesophageal reflux disease; no. of active problems of congestive heart failure; no. of active problems of diabetes; no. of active problems of dyspnea; no. of active problems of hypertension; no. of active problems of obesity; no. of active problems of rhinitis; no. of active problems of sleep apnea; no. of active problems about smoking; no. of active problems of wheezing; and the priority of the last active problem of COPD.	Number of features: 17.
Medication	Total no. of COPD medications ordered; no. of COPD medication orders; total no. of distinct COPD medications ordered; total no. of COPD medication refills allowed; total no. of units of COPD medications ordered; no. of COPD reliever orders; total no. of medications in COPD reliever orders; total no. of distinct medications in COPD reliever orders; total no. of refills allowed for COPD relievers; total no. of units of COPD relievers ordered; no. of COPD controller orders; total no. of medications in COPD controller orders; total no. of distinct medications in COPD controller orders; total no. of refills allowed for COPD controllers; total no. of units of COPD controllers ordered; whether a nebulizer was used; no. of nebulizer medication orders; total no. of medications in nebulizer medication orders; total no. of distinct medications in nebulizer medication orders; total no. of refills allowed for nebulizer medications; total no. of units of nebulizer medications ordered; whether a spacer was used; no. of medication orders; total no. of medications ordered; total no. of distinct medications ordered; total no. of units of medications ordered; total no. of medication refills allowed; total no. of short-acting muscarinic antagonists (SAMA) ordered; total no. of refills allowed for SAMA; total no. of units of SAMA ordered; total no. of inhaled corticosteroids (ICS) ordered; total no. of refills allowed for ICS; total no. of units of ICS ordered; total no. of short-acting beta-2 agonists (SABA) ordered; total no. of refills allowed for SABA; total no. of units of SABA ordered; total no. of	COPD medication categories [3,112]: • Short-term relievers: systemic corticosteroid [50]; short-acting muscarinic antagonist (SAMA) [50]; short-acting beta-2 agonist (SABA); SABA and SAMA combination; and mucolytic agent. • Long-term controllers: inhaled corticosteroid (ICS); long-acting muscarinic antagonist (LAMA); long-acting beta-2 agonist (LABA); LABA and LAMA combination; ICS and LABA combination; ICS, LABA, and LAMA combination; methylxanthine; anti-interleukin-5; anti-interleukin-5 receptor; and phosphodiesterase-4 inhibitor (PDE-4). Number of features: 52.

	systemic corticosteroid ordered; total no. of refills allowed for systemic corticosteroids; total no. of units of systemic corticosteroids ordered; total no. of long-acting beta-2 agonists (LABA) ordered; total no. of refills allowed for LABA; total no. of units of LABA ordered; total no. of long-acting muscarinic antagonists (LAMA) ordered; total no. of refills allowed for LAMA; total no. of units of LAMA ordered; total no. of phosphodiesterase-4 inhibitors (PDE-4) ordered; total no. of refills allowed for PDE-4; total no. of units of PDE-4 ordered; total no. of ICS and LABA combinations ordered; total no. of ICS, LABA, and LAMA combinations ordered; total no. of LABA and LAMA combinations ordered; and total no. of SABA and SAMA combinations ordered.	
Insurance	Computed based on the end of the index year: whether the patient had any private insurance; whether the patient had any public insurance; and whether the patient was paid by oneself or a charity.	Number of features: 3.
Encounter	No. of all types of encounters; no. of major encounters for COPD [28,50]; no. of outpatient visits; no. of outpatient visits with a primary diagnosis of COPD; no. of emergency department (ED) visits; average length of stay of an ED visit; no. of ED visits related to COPD [28]; no. of inpatient stays; total length of inpatient stays; average length of an inpatient stay; no. of inpatient stays related to acute COPD exacerbation or respiratory failure; no. of encounters related to acute COPD exacerbation or respiratory failure [18,60]; no. of outpatient visits to the patient's primary care provider (PCP); no. of admissions to the intensive care unit; admission type of the most emergent encounter (elective, urgent, emergency, or trauma); admission type of the last encounter (elective, urgent, emergency, or trauma); type of the last encounter (ED visit, outpatient visit, or inpatient stay); type of the first encounter (ED visit, outpatient visit, or inpatient stay) related to COPD in the data set; length of stay of the last ED visit; no. of ED visits in the past 6 months; no. of inpatient stays in the past 6 months; and no. of major encounters for COPD in the past 6 months.	A major encounter for COPD was defined as an ED visit having a COPD diagnosis code, an inpatient stay having a COPD diagnosis code, or an outpatient visit having a primary diagnosis of COPD. All else being equal and compared with a patient with only outpatient visits with COPD as a secondary diagnosis, a patient with ≥1 major encounter for COPD is more likely to have severe COPD exacerbations in the future. Number of features: 22.
Visit status and appointment scheduling	The day of the week when the last ED visit started; the last encounter's discharge disposition location (home, left against medical advice, or other non-home location); no. of times of leaving against medical advice; no. of no shows; no. of cancelled appointments; no. of days since the last inpatient stay; no. of days since the last outpatient visit; no. of days since the last outpatient visit on COPD; no. of days since the last ED visit; no. of days since the last ED visit on COPD [62]; the shortest time between making the request and the actual visit among all occurred encounters; no. of days between making the request and the actual visit of the last encounter; no. of visits having same day appointments; and whether the last inpatient stay came from the ED.	Number of features: 14.
Patient's care continuity degree	No. of distinct medication prescribers; no. of distinct COPD medication prescribers; no. of distinct providers seen in outpatient visits; no. of distinct PCPs of the patient; and no. of distinct ED locations the patient went to (including inpatient stays admitted from the ED).	Number of features: 5.
Procedure	No. of CPT and HCPCS procedure codes; no. of ICD-10 and ICD-9 procedure codes; no. of CPT procedure codes of the fractional exhaled nitric oxide test; no. of CPT procedure codes of spirometry; no. of HCPCS procedure codes of home oxygen therapy; no. of CPT and HCPCS procedure codes of influenza vaccination; and whether mechanical ventilation was recorded using ICD-10 and ICD-9 procedure codes.	Number of features: 7.
Allergy	No. of the patient's allergies; indicator of food allergy; indicator of drug or material allergy; indicator of environmental allergy; the greatest severity of the patient's food allergies; the greatest severity of the	Number of features: 7.

	patient's drug or material allergies; and the greatest severity of the patient's environmental allergies.	
Social behavior history	Whether the patient was last recorded as a current smoker; whether the patient was last recorded as a former smoker; the last recorded no. of packs of cigarettes the patient consumed per day; the average no. of packs of cigarettes the patient consumed per day across all of the records; no. of years the patient had smoked for based on the last record; whether the patient was ever documented of consuming alcohol; whether the patient consumed alcohol based on the last record; the last recorded no. of fluid ounces of alcohol the patient consumed per week; the average no. of fluid ounces of alcohol the patient consumed per week across all of the records; the last recorded no. of alcohol drinks the patient consumed per week; the average no. of alcohol drinks the patient consumed per week across all of the records; whether the patient took any illicit drug based on the last record; whether the patient was ever documented of taking any illicit drug; the last recorded no. of times the patient took illicit drugs per week; and the average no. of times the patient took illicit drugs per week across all of the records.	Number of features: 15.
Provider	The PCP's type (physician, nurse, physician assistant, or other); whether the PCP is a resident; the PCP's clinician title (Doctor of Medicine, registered nurse, physician assistant, or other); the PCP's age; whether the patient and the PCP are of the same gender; no. of years for which the PCP had practiced at the UWM; no. of patients with COPD of the PCP; and the percentage of the PCP's patients with COPD in the pre-index year having severe COPD exacerbations in the index year.	A patient's current PCP was defined as the patient's PCP known at the last outpatient visit. Number of features: 8.

Table 2. The features employed in the final model in the main analysis and their importance values.

Rank	Feature	Importance value in % based on the feature's apportioned contribution to the model
1	Number of days since the last diagnosis code of acute COPD exacerbation	8.384
2	Number of COPD diagnoses	5.287
3	Number of diagnoses of acute COPD exacerbation	5.107
4	Number of days since the last ED visit	4.851
5	Number of major encounters	4.583
6	Number of primary or principal COPD diagnoses	3.571
7	Number of days since the last COPD diagnosis code	2.249
8	Average SpO_2	2.016
9	Maximum body mass index	2.003
10	Number of ED visits	1.947
11	Number of ED visits in the past 6 months	1.900
12	Number of years from the first encounter related to COPD	1.846
13	Whether the patient is married	1.659
14	Number of days since the last ED visit related to COPD	1.534
15	Average length of stay of an ED visit	1.521
16	Average respiratory rate	1.275
17	Number of CPT procedure codes	1.128
18	Average heart rate	1.056
19	Total number of distinct medications ordered	1.013
20	Number of no shows	0.994
21	Minimum SpO_2	0.990
22	Average temperature	0.965
23	Whether the first COPD diagnosis was given at an outpatient visit	0.951
24	Number of encounters related to acute COPD exacerbation or respiratory failure	0.928

25	Number of days between making the request and the actual visit of the last encounter	0.864
26	Whether the patient is Hispanic	0.834
27	Number of days since the last laboratory test	0.802
28	Whether the admission type of the most emergent encounter was elective	0.758
29	Number of medication orders	0.748
30	Total number of distinct medications in COPD reliever orders	0.708
31	Whether the admission type of the most emergent encounter was emergency	0.703
32	Relative change of body mass index	0.661
33	Number of ICD-9 and ICD-10 diagnosis codes	0.658
34	Number of days since the last outpatient visit	0.656
35	Maximum heart rate	0.649
36	Day of the week when the last ED visit started	0.631
37	Breathing abnormality like dyspnea	0.582
38	Relative change of weight	0.580
39	Total number of COPD medication refills allowed	0.557
40	Maximum red blood cell count	0.548
41	Maximum temperature	0.546
42	Maximum height	0.540
43	Maximum Hgb level	0.537
44	Maximum white blood cell count	0.534
45	Total number of units of long-acting muscarinic antagonists ordered	0.530
46	Average diastolic blood pressure	0.516
47	Total number of distinct medications in COPD medication orders	0.499
48	Average systolic blood pressure	0.493
49	Age	0.468
50	Maximum eosinophil count	0.467
51	Total number of units of medications ordered	0.466
52	Average length of an inpatient stay	0.463
53	Total number of medications ordered	0.460
54	The shortest time between making the request and the actual visit among all occurred encounters	0.457
55	Number of days since the last outpatient visit on COPD	0.454
56	Number of COPD medication orders	0.439
57	Number of nebulizer medication orders	0.434
58	Maximum neutrophil count	0.434
59	Whether the patient is a black or an African American	0.421
60	Maximum percentage of eosinophils	0.419
61	Number of major encounters in the past 6 months	0.414
62	Maximum systolic blood pressure	0.400
63	Number of years the patient had smoked for based on the last record	0.399
64	Number of laboratory tests	0.379
65	Maximum percentage of neutrophils	0.374
66	Average number of packs of cigarettes the patient smoked per day across all of the records	0.368
67	Number of ED visits related to COPD	0.364
68	Number of active problems	0.360
69	Length of stay of the last ED visit	0.354
70	Total number of medication refills allowed	0.354
71	Maximum respiratory rate	0.349
72	Maximum diastolic blood pressure	0.341
73	Total number of units of COPD controllers ordered	0.335
74	Minimum Hgb level	0.332
75	Total length of inpatient stays	0.322
76	Number of outpatient visits to the PCP	0.318
77	Maximum PaO_2	0.312

78	Minimum Hct	0.301
79	Total number of refills allowed for COPD controllers	0.299
80	Total number of medications in nebulizer medication orders	0.298
81	Total number of short-acting beta-2 agonists ordered	0.296
82	Whether the first COPD diagnosis was given at an ED visit	0.295
83	Number of CPT and HCPCS procedure codes of influenza vaccination	0.291
84	Number of laboratory tests with abnormal results	0.288
85	The last recorded number of packs of cigarettes the patient consumed per day	0.288
86	Whether the patient was last recorded as a current smoker	0.283
87	Total number of long-acting muscarinic antagonists ordered	0.269
88	Number of distinct medication prescribers	0.268
89	Total number of long-acting beta-2 agonists ordered	0.260
90	Number of cancelled appointments	0.259
91	Whether the patient is a white	0.257
92	Maximum Hct	0.249
93	Total number of units of COPD medications ordered	0.223
94	Maximum HbA1c	0.223
95	Total number of refills allowed for long-acting muscarinic antagonists	0.223
96	Total number of COPD medications ordered	0.222
97	Total number of refills allowed for inhaled corticosteroids	0.219
98	Number of same day appointments	0.213
99	Priority of the last active problem of COPD	0.212
100	Number of all types of encounters	0.211
101	Total number of units of COPD relievers ordered	0.211
102	Total number of refills allowed for long-acting beta-2 agonists	0.209
103	Whether the patient's PCP is a Doctor of Medicine	0.208
104	Total number of inhaled corticosteroid and long-acting beta-2 agonist combinations ordered	0.206
105	Number of COPD controller orders	0.195
106	Number of COPD reliever orders	0.194
107	Total number of distinct medications in nebulizer medication orders	0.187
108	Number of distinct ED locations the patient went to (including inpatient stays admitted from the ED)	0.187
109	Maximum CRP	0.185
110	Whether the last encounter was an ED visit	0.185
111	Total number of inhaled corticosteroids ordered	0.184
112	Number of ICD-10 and ICD-9 procedure codes	0.172
113	Total number of units of short-acting muscarinic antagonists ordered	0.172
114	Total number of units of nebulizer medications ordered	0.172
115	Minimum $PaCO_2$	0.172
116	Whether the last encounter's discharge disposition location was home	0.170
117	Total number of refills allowed for short-acting beta-2 agonists	0.169
118	Total number of short-acting muscarinic antagonists ordered	0.165
119	Number of active problems about smoking	0.163
120	Number of days since the last inpatient stay	0.162
121	Number of distinct COPD medication prescribers	0.161
122	Total number of distinct medications in COPD controller orders	0.148
123	The greatest severity of the patient's drug or material allergies	0.143
124	Total number of units of inhaled corticosteroids ordered	0.143
125	Number of active problems of COPD	0.140
126	Total number of short-acting beta-2 agonist and short-acting muscarinic antagonist combinations ordered	0.137
127	Total number of units of short-acting beta-2 agonists ordered	0.132
128	Total number of systemic corticosteroids ordered	0.128

129	Number of outpatient visits	0.126
130	Whether the last COPD diagnosis was a primary or principal diagnosis	0.121
131	Number of patients with COPD of the PCP	0.120
132	Average number of fluid ounces of alcohol the patient consumed per week across all of the records	0.116
133	Number of allergies	0.116
134	Age of the PCP	0.113
135	Total number of medications in COPD controller orders	0.109
136	Total number of medications in COPD reliever orders	0.109
137	The last recorded number of alcohol drinks the patient consumed per week	0.108
138	Whether the patient was paid by oneself or a charity on the last day	0.107
139	The last recorded number of fluid ounces of alcohol the patient consumed per week	0.106
140	Whether the patient is divorced	0.095
141	Whether the patient was last recorded as a former smoker	0.093
142	Whether the patient is widowed	0.093
143	Average number of alcohol drinks the patient consumed per week across all of the records	0.092
144	Minimum SaO_2	0.087
145	The percentage of the PCP's patients with COPD in the pre-index year having severe COPD exacerbations in the index year	0.087
146	Whether the patient's PCP is a physician	0.087
147	Whether the patient had any private insurance on the last day	0.085
148	Number of CPT procedure codes of spirometry	0.084
149	Total number of refills allowed for COPD relievers	0.082
150	Total number of units of long-acting beta-2 agonists ordered	0.080
151	Number of active problems of hypertension	0.079
152	Total number of refills allowed for short-acting muscarinic antagonists	0.078
153	Whether the patient took any illicit drug based on the last record	0.078
154	Number of inpatient stays related to acute COPD exacerbation or respiratory failure	0.076
155	Whether the patient is single	0.074
156	Minimum PaO_2	0.074
157	Number of active problems of sleep apnea	0.073
158	Whether the patient had any public insurance on the last day	0.072
159	Peripheral vascular disease	0.071
160	Number of distinct providers seen in outpatient visits	0.070
161	Upper respiratory tract infection	0.068
162	Substance use	0.065
163	Whether a nebulizer was used	0.065
164	Whether the patient has any material or drug allergy	0.061
165	Whether the patient used a spacer	0.059
166	Number of distinct PCPs of the patient	0.058
167	Maximum $PaCO_2$	0.057
168	Number of outpatient visits with a primary diagnosis of COPD	0.057
169	Total number of refills allowed for nebulizer medications	0.057
170	Number of inpatient stays	0.054
171	Congestive heart failure	0.054
172	Whether the patient consumed alcohol based on the last record	0.054
173	Dementia	0.051
174	Whether the patient is an Asian	0.049
175	Total number of units of systemic corticosteroids ordered	0.048
176	Mental disorder	0.046
177	Number of inpatient stays in the past 6 months	0.045
178	Whether the patient was ever documented of consuming alcohol	0.043
179	Cerebrovascular disease	0.041
180	Whether the patient was ever documented of taking any illicit drug	0.039

181	Metastatic solid tumor	0.038
182	Whether the last inpatient stay came from the ED	0.038
183	Number of active problems of diabetes	0.036
184	Number of active problems of anxiety/depression	0.034
185	Hypertension	0.033
186	Renal disease	0.033
187	Number of active problems of acute COPD exacerbation	0.030
188	Obesity	0.029
189	Whether the maximum Hgb was abnormally high	0.027
190	Number of active problems of gastroesophageal reflux disease	0.027
191	Gastroesophageal reflux	0.027
192	Asthma	0.026
193	Whether the patient speaks English	0.025
194	Myocardial infarction	0.023
195	Total number of refills allowed for systemic corticosteroids	0.022
196	Whether the patient is a female	0.020
197	Whether the patient is separated from the spouse	0.020
198	Diabetes without chronic complication	0.020
199	Whether the last encounter was an inpatient stay	0.020
200	Whether the minimum Hct was abnormally low	0.019
201	Whether the admission type of the last encounter was elective	0.018
202	Whether the maximum Hct was abnormally high	0.017
203	Average peak expiratory flow rate	0.016
204	Whether the patient had undergone mechanical ventilation	0.016
205	Lung cancer	0.016
206	Whether the patient is a Native Hawaiian or an other Pacific Islander	0.015
207	Mild liver disease	0.014
208	Cirrhosis	0.014
209	Hemiplegia or paraplegia	0.014
210	Number of diagnoses of noncompliance with medication regimen	0.013
211	Pneumonia	0.013
212	Whether the patient's PCP's title is nurse	0.013
213	Ischemic heart disease	0.013
214	Acquired immunodeficiency syndrome	0.013
215	Whether the patient has any environmental allergy	0.011
216	Vitamin D deficiency	0.011
217	Number of active problems of congestive heart failure	0.010
218	The last recorded number of times the patient took illicit drugs per week	0.010
219	Malignancy	0.008
220	Number of active problems of dyspnea	0.008
221	Number of active problems of asthma	0.008
222	Gastrointestinal bleeding	0.008
223	Average number of times the patient took illicit drugs per week across all of the records	0.008
224	Cataract	0.007
225	Diabetes with chronic complication	0.006
226	Number of active problems of rhinitis	0.006
227	Whether the minimum Hgb was abnormally low	0.006
228	Number of active problems of obesity	0.005
229	Allergic rhinitis	0.004

We used the xgb.save() function in the xgboost package of R to save our final XGBoost model to a file in binary format. This file is available at http://faculty.washington.edu/luogang/COPD_care_model_UW.

Abbreviations:

A1AT: Alpha-1 antitrypsin
BMI: body mass index
COPD: chronic obstructive pulmonary disease
CPT: Current Procedural Terminology
CRP: C-reactive protein
ED: emergency department
FEV_1: forced expiratory volume in 1 second
FVC: forced vital capacity
HCPCS: Healthcare Common Procedure Coding System
Hct: hematocrit
HbA1c: hemoglobin A1c
Hgb: hemoglobin
ICD-10: International Classification of Diseases, Tenth Revision
ICD-9: International Classification of Diseases, Ninth Revision
ICS: inhaled corticosteroid
IgA: immunoglobulin A
IgE: immunoglobulin E
LABA: long-acting beta-2 agonist
LAMA: long-acting muscarinic antagonist
$PaCO_2$: arterial partial pressure of carbon dioxide
PaO_2: arterial partial pressure of oxygen
PCP: primary care provider
PDE-4: phosphodiesterase-4 inhibitor
SABA: short-acting beta-2 agonist
SAMA: short-acting muscarinic antagonist
SaO_2: arterial oxygen saturation
SpO_2: peripheral capillary oxygen saturation

References

102. Guinness World Records. The world's oldest people and their secrets to a long life. 2020. https://www.guinnessworldrecords.com/news/2020/10/the-worlds-oldest-people-and-their-secrets-to-a-long-life-632895.
103. Guinness World Records. Lightest birth. 2020. https://www.guinnessworldrecords.com/world-records/lightest-birth.
104. Guinness World Records. Heaviest man ever. 2020. https://www.guinnessworldrecords.com/world-records/heaviest-man.
105. Guinness World Records. Shortest baby. 2020. https://www.guinnessworldrecords.com/world-records/shortest-baby.
106. Guinness World Records. Tallest man ever. 2020. https://www.guinnessworldrecords.com/world-records/tallest-man-ever.
107. Gwyneth O. Part V Fat: no more fear, no more contempt. The Eating Disorder Institute. 2011. https://edinstitute.org/blog/2011/12/8/part-v-fat-no-more-fear-no-more-contempt.
108. Wikipedia. List of heaviest people. 2021. https://en.wikipedia.org/w/index.php?title=List_of_heaviest_people&oldid=1000662342.
109. Hankinson JL, Odencrantz JR, Fedan KB. Spirometric reference values from a sample of the general U.S. population. Am J Respir Crit Care Med 1999 Jan;159(1):179-187. PMID:9872837
110. Pellegrino R, Viegi G, Brusasco V, Crapo RO, Burgos F, Casaburi R, Coates A, van der Grinten CP, Gustafsson P, Hankinson J, Jensen R, Johnson DC, MacIntyre N, McKay R, Miller MR, Navajas D, Pedersen OF, Wanger J. Interpretative strategies for lung function tests. Eur Respir J 2005 Nov;26(5):948-968. PMID:16264058
111. Marion MS, Leonardson GR, Rhoades ER, Welty TK, Enright PL. Spirometry reference values for American Indian adults: results from the Strong Heart Study. Chest 2001 Aug;120(2):489-495. PMID:11502648
112. National Jewish Health. Bronchodilators. 2018. https://nationaljewish.org/conditions/medications/copd/bronchodilators.