Chronic Obstructive Pulmonary Diseases:
Journal of the COPD Foundation®

Original Research

Identification of Severe Acute Exacerbations of Chronic Obstructive Pulmonary Disease Subgroups by Machine Learning Implementation in Electronic Health Records

Huan Li, MS[1] John Huston, MD[1] Jana Zielonka, MD[1] Shannon Kay, MD, MS[1] Maor Sauler, MD[1] Jose Gomez, MD, MS[1,2]

Abstract

Rationale: Acute exacerbations of chronic obstructive pulmonary disease (AECOPDs) are heterogeneous. Machine learning (ML) has previously been used to dissect some of the heterogeneity in COPD. The widespread adoption of electronic health records (EHRs) has led to the rapid accumulation of large amounts of patient data as part of routine clinical care. However, it is unclear whether the implementation of ML in EHR-derived data has the potential to identify subgroups of AECOPD.

Objectives: To determine whether ML implementation using EHR data from severe AECOPDs requiring hospitalization identifies relevant subgroups.

Methods: This study used 2 retrospective cohorts of patients with AECOPDs (non-COVID-19 and COVID-19) treated at Yale-New Haven Hospital. K-means clustering was used to identify patient subgroups.

Measurements and Main Results: We identified 3 subgroups in the non-COVID cohort (n=1736). Each subgroup had distinct clinical characteristics. The reference subgroup was the largest (n=904), followed by cardio-renal (n=548) and eosinophilic (n=284). The eosinophilic subgroup had milder severity of AECOPD, including a shorter hospital stay ($p<0.01$). The cardio-renal subgroup had the highest mortality during (5%) and in the year after hospitalization (30%). Validation of the severe AECOPD classifier in the COVID-19 cohort recapitulated the characteristics seen in the non-COVID cohort. AECOPD subgroups in the COVID-19 cohort had different interleukin (IL)-1 beta, IL-2R, and IL-8 levels (false discovery rate ≤0.05). These specific leukocyte and cytokine profiles resulted in inflammatory differences between the AECOPD subgroups based on C-reactive protein levels.

Conclusions: Incorporating ML with EHR data allows the identification of specific clinical and biological subgroups for severe AECOPD.

1. Pulmonary, Critical Care and Sleep Medicine Section, Yale University, New Haven, Connecticut, United States
2. Center for Precision Pulmonary Medicine, Yale University, New Haven, Connecticut, United States

Abbreviations:

ABG=arterial blood gas; **AECOPD**=acute exacerbations of COPD; **AIC**=Akaike information criterion; **BMI**=body mass index; **BUN**=blood urea nitrogen; **CRP**=C-reactive protein; **COPD**=chronic obstructive pulmonary disease; **EHR**=electronic health record; **FDR**=false discovery rate; **HF**=heart failure; **HR**=hazard ratio; **ICS/LABA**=inhaled corticosteroid/long-acting beta2-agonist combination; **ICU**=intensive care unit; **IL**=interleukin; **ML**=machine learning; **pro-BNP**=probrain natriuretic peptide; **T2**=type 2

Funding Support:

R01 HL153604, and R03 HL154275 to JLG. This publication was made possible by CTSA Grant Number UL1 TR000142 from the National Center for Advancing Translational Science (NCATS), a component of the National Institutes of Health (NIH). National Heart, Lung, and Blood Institute, 2T32HL007778-26 to support JZ and SK. R01 HL155948 to MS. Manuscript contents are solely the responsibility of the authors and do not necessarily represent the official view of the NIH.

Citation:

Li H, Huston J, Zielonka J, Kay S, Sauler M, Gomez J. Identification of severe acute exacerbations of chronic obstructive pulmonary disease subgroups by machine learning implementation in electronic health records. *Chronic Obstr Pulm Dis.* 2024;11(6):611-623. doi: https://doi.org/10.15326/jcopdf.2024.0556

Publication Dates:

Date of Acceptance: October 13, 2024
Published Online Date: October 17, 2024

Address correspondence to:

Jose Gomez, MD
Yale University
300 Cedar Street (S419 TAC)
New Haven, CT 06520-8057
Phone: (203) 785-4163
Email: jose.gomez-villalobos@yale.edu

Keywords:

electronic health records; machine learning; acute exacerbations of COPD; clusters

This article has an online supplement.

Introduction

Chronic obstructive pulmonary disease (COPD) is heterogeneous.[1-3] Factors involved in heterogeneity include clinical characteristics, distinct pathobiological characteristics, including types of inflammation, genetic factors, and treatment response. The emergence of the concept of endotypes has led to the development of novel disease classification models.[4] Acute exacerbations of COPD (AECOPDs) also exhibit this heterogeneity, which can be related to the baseline characteristics of the subgroups or to the triggers for exacerbations.[5]

Severe AECOPDs that require hospitalization are associated with significant morbidity and mortality, in addition to significant health care expenses.[6] Furthermore, the Centers for Medicare and Medicaid Services has established a Hospital Readmission Reduction Program that penalizes hospitals that have high readmission rates for COPD.[7] All these factors underscore the impact of severe AECOPDs on patients and the health care system and underscore the importance of understanding the heterogeneity associated with severe exacerbations.

Several studies have shown the ability of machine learning (ML) methods to identify discrete groups in COPD.

COPD subgroups have been identified by: (1) cytokine profiles[2]; (2) a combination of clinical data including comorbidities[8]; (3) a combination of clinical, physiologic, and imaging data[1]; and (4) imaging,[9] among others. A new eosinophilic endotype of COPD has also been identified thanks to advances in our understanding of COPD pathobiology.[10,11] Despite these important observations, the highly selected cohorts used to obtain these insights may not reflect the overall COPD patient population.

The 2009 Federal Health Information Technology for Economic and Clinical Health Act led to the creation of an incentive program to encourage hospitals and health care providers to adopt electronic health records (EHRs). Currently, more than 95% of U.S. hospitals have adopted EHRs.[12] As a result of EHR adoption, the volume of health care data has increased exponentially[13] from 153 exabytes in 2013 to 2314 exabytes in 2020. This massive increase in data encodes millions of health care encounters and creates a crucial opportunity to transform patient care. The concept of computable phenotypes, defined as clinical conditions or characteristics that are derived from a computerized query using a defined set of data elements,[14] has gained significant attention as a result. By leveraging EHR data, clinical decision-making in COPD can be informed by novel computational applications.

Identifying disease subgroups and potential disease endotypes using EHR data may help focus therapeutic efforts on COPD exacerbations. The purpose of this study was to determine whether the combination of EHR data and ML in hospitalizations for severe AECOPDs could identify specific subgroups of patients characterized by differences in clinical outcomes.

Methods

Original Cohort Data Source and Study Population

We conducted a retrospective cohort study using data collected from patients hospitalized at Yale-New Haven Hospital between September 30, 2012, after the Epic EHR system (Epic; Verona, Wisconsin) was implemented, and December 31, 2017. The Yale University Human Research Protection Program approved this study and ethical approval was obtained from the Yale Institutional Review Board under a Waiver of Consent. We have previously described this cohort.[15] Data were obtained from the Joint Data Analytics Team at Yale University School of Medicine.

COVID Cohort

The Yale Department of Medicine COVID-19 Explorer and Repository tool was used to extract data on patients admitted with COVID-19 from March 1, 2020, to April 1, 2021, in Yale-New Haven Health System hospitals.[16] The patients had a positive test for SARS-CoV-2 using reverse transcriptase–polymerase chain reaction assays performed on nasopharyngeal swab specimens within 14 days after admission.

Clustering

To use the unsupervised learning k-means clustering method, we preprocessed the non-COVID-19 data. We identified those features with missing values and removed them to ensure that the training process was unbiased and free of unnecessary noise. This led to a data frame with 1736 observations and 52 features, including the unique identifier. We did not use imputation for the selected features and only used complete data. The numerical features (24) were normalized, while the string features (27) were one hot encoded. We utilized an autoencoding deep learning technique to enhance the efficiency of k-means clustering on datasets by reducing the datasets' dimensions to 3. Prior to training the k-means clustering model, we employed the NbClust Package in R to determine the optimal number of clusters. Once the number of clusters was identified, we divided the data into 80% for training and 20% for testing purposes.

Classifier

An XGBoost classifier was developed using the multi:softmax objective function to target the subgroup labels obtained from the previous k-means clustering. The same data processing methods were applied, and the data was divided into 80% for training and 20% for testing. A Grid Search was conducted with 5-fold cross-validation to identify the best hyperparameters for the classifier. The trained classifier was then saved and later applied to the COVID-19 cohort. The classifier code is included in the supplementary material in the online supplement.

Statistical Analysis

The R statistical software was used for statistical analyses. Significance was defined as $p<0.05$ and false discovery rate (FDR) <0.05. STROBE guidelines for cohort studies were followed in the preparation of this report. Additional methods are described in the supplementary material in the online supplement.

Results

Identification of the COPD Subgroups

To identify subgroups characterized by specific clinical features, we applied k-means, an unsupervised clustering method, to clinical data from 1736 patients admitted to the hospital for a severe AECOPD. We used 51 features to implement this clustering method. The resulting subgroups were characterized by clinical similarities. We identified 3 distinct subgroups in the resulting analysis (Table 1). Across all 3 subgroups, sex and absolute monocyte counts were similar, suggesting that sex or monocytes were not key factors in this classification.

Clinical Characteristics of the Acute Exacerbation of COPD Subgroups

The largest subgroup (n=904, 52%) was mainly composed of former smokers (69%), with the highest rates of comorbid hypertension of all subgroups (94%). Half of these patients were diagnosed with heart failure (50%) or diabetes (54%). This subgroup was also characterized by the highest inpatient administration of inhaled corticosteroid/long-acting beta2-agonist combination (ICS/LABA), antibiotics, and systemic steroids. As the most prevalent subgroup, it will be treated as a reference herein.

The patients in the second largest subgroup (n=548, 32%) were the oldest (77 years [70–87]) and had the lowest body mass index (BMI) of the 3 subgroups (25.6 kg/m^2 [21.8–31.1]). This subgroup was notable for the highest rates of heart failure (62%) and chronic kidney disease (42%). This subgroup had the lowest systemic steroid administration rate (73%) and ICS/LABA (53%) of the 3 subgroups but had similar rates of antibiotic use to the reference subgroup (87%). Given the high rates of heart failure and renal failure, this subgroup will be described as cardio-renal hereafter.

The third and smallest subgroup (n=284, 16%) had the youngest patients (61 years [54–72]) and the highest rate of active smokers (52%). Subgroup 3 had the lowest rates of heart failure (38%) and chronic kidney disease (23%), but the highest rates of allergic rhinitis (12%) in the 3 subgroups. This subgroup also had the lowest antibiotic administration rates (77%). Consistent with the high rates of active smoking, subgroup 3 had the highest rate of nicotine replacement during hospitalization (44%).

Subgroups of Acute Exacerbation of COPD Exhibit Distinct Blood Chemistry and Complete Blood Counts

Although blood chemistries were not used to identify the COPD subgroups, we were interested in exploring whether the cardio-renal subgroup also showed abnormal markers of cardiac and renal function. We compared the values of probrain natriuretic peptide (pro-BNP), blood urea nitrogen (BUN), and creatinine values from patients in the 3 subgroups. The cardio-renal subgroup had the highest combined pro-BNP, BUN, and creatinine values of the 3 subgroups (Figure 1A-C).

In contrast to blood chemistries, complete blood count values were used to identify COPD subgroups. Consequently, white blood cell, neutrophil, lymphocyte, basophil, and eosinophil counts significantly differed among the subgroups (Table 1). Subgroup 3 was characterized by the lowest neutrophil counts (5400 cells/microliter [4000–7300]), and highest blood lymphocyte (2325 cells/microliter [1637–3,039]) and eosinophil counts (337 cells/microliter [96–396]) (Figure 1 D-F). Due to the increasing recognition that eosinophils are a major risk factor for COPD exacerbations,[17-19] the identification of a subgroup with higher counts is particularly relevant. Subgroup 3 will be described as eosinophilic hereafter.

COPD Subgroups are Characterized by Specific Disease Outcomes

Given the known associations between specific comorbidity patterns,[20] eosinophilic inflammation in COPD exacerbations,[17] and exacerbation outcomes, we examined whether the COPD exacerbation subgroups demonstrated any outcome differences. We found no differences in intensive care use or readmissions within 30 days. Consistent with previous observations,[17] we found that the eosinophilic subgroup had the shortest stay (5.98 days [2–6]) (Table 2). During hospitalization (5%) and in the year following an AECOPD hospitalization (30%), the cardio-renal subgroup had the highest mortality rates.

Table 1. Clinical Characteristics of COPD Subgroups

	Reference (n=904) 52%	Cardio-Renal (n=548) 32%	Eosinophilic (n=284) 16%	*P*-value
Age (years)	71 (61–79)	79 (70–87)	61 (54–72)	<0.001
Female Sex n (%)	504 (55.8)	306 (55.8)	153 (53.9)	0.84
Race n (%)				<0.01
White	715 (79)	478 (87)	210 (74)	
Black	131 (15)	41 (8)	49 (17)	
Other	51 (6)	25 (5)	24 (8)	
Hispanic Ethnicity n (%)	52 (6)	18 (3)	28 (10)	<0.01
BMI (kg/m^2)	27.8 (23.0–34.1)	25.6 (21.8–31.10)	28.4 (23.4–34.8)	<0.001
Smoking Status (n)				<0.001
Never n (%)	72 (8)	47 (9)	11 (4)	
Current n (%)	210 (23)	94 (17)	149 (52)	
Former n (%)	622 (69)	407 (74)	124 (44)	
Comorbidities				
Heart Failure n (%)	448 (50)	341 (62)	107 (38)	<0.001
Cerebrovascular Disease n (%)	197 (22)	85 (16)	58 (20)	0.01
Diabetes Mellitus n (%)	491 (54)	207 (38)	151 (53)	<0.001
Chronic Kidney Disease n (%)	354 (39)	232 (42)	66 (23)	<0.001
Allergic Rhinitis n (%)	69 (8)	27 (5)	33 (12)	0.002
Lung Cancer n (%)	134 (15)	103 (19)	33 (12)	0.02
Sleep Apnea n (%)	346 (38)	116 (21)	109 (38)	<0.001
Gastroesophageal Reflux n (%)	554 (61)	225 (41)	147 (52)	<0.001
Hypertension n (%)	846 (94)	460 (84)	231 (81)	<0.001
Multiple Comorbidities n (%)	836 (92)	477 (87)	246 (87)	<0.001
Medications				
Albuterol n (%)	631 (70)	351 (64)	199 (70)	0.05
Antibiotic n (%)	814 (90)	474 (87)	219 (77)	<0.001
Inhaled Corticosteroids n (%)	79 (9)	76 (14)	29 (10)	0.009
Inhaled Corticosteroid with Long-Acting Beta2-Agonist n (%)	640 (71)	293 (53)	185 (65)	<0.001
Long-Acting Muscarinic Antagonist n (%)	429 (47)	210 (38)	132 (48)	0.002
Leukotriene Receptor Antagonist n (%)	123 (14)	46 (8)	47 (17)	0.001
Nicotine Replacement n (%)	208 (23)	55 (10)	124 (44)	<0.001
Systemic Steroids n (%)	787 (87)	398 (73)	225 (79)	<0.001
Complete Blood Count				
White Blood Cells (10^3 cells/μL)	10.0 (7.5–12.9)	10.0 (7.3–13.1)	9.1 (7.1–11.5)	0.008
Absolute Neutrophil Count (cells/μL)	7.6 (5.3–10.6)	7.9 (5.4–11.2)	5.4 (4.0–7.3)	<0.001
Absolute Eosinophil Count (cells/μL)	78 (0–174.3)	0 (0–151.7)	237 (96.0–396.4)	<0.001
Absolute Basophil Count (cells/μL)	0 (0–42.7)	0 (0–33)	73.2 (28.5–108.0)	<0.001
Absolute Monocyte Count (cells/μL)	706.5 (489.6–963.0)	736.0 (492.6–1002.0)	760.0 (569.5–960.0)	0.07
Absolute Lymphocyte Count (cells/μL)	1158.7 (767.8–1677.8)	990.8 (681.5–1452.0)	2325.0 (1636.0–3039.0)	<0.001
Hematocrit (%)	38.4 (34.3–42.6)	37.4 (32.8–41.4)	41.7 (38.0–44.8)	<0.001
Hemoglobin (g/dL)	12.6 (11.1–14.1)	12.3 (10.7–13.5)	13.8 (12.5–15.0)	<0.001
Platelets (10^3 cells/μL)	228.0 (177.8–294.3)	211.5 (164.0–273.3)	238.5 (179.8–297.2)	<0.001

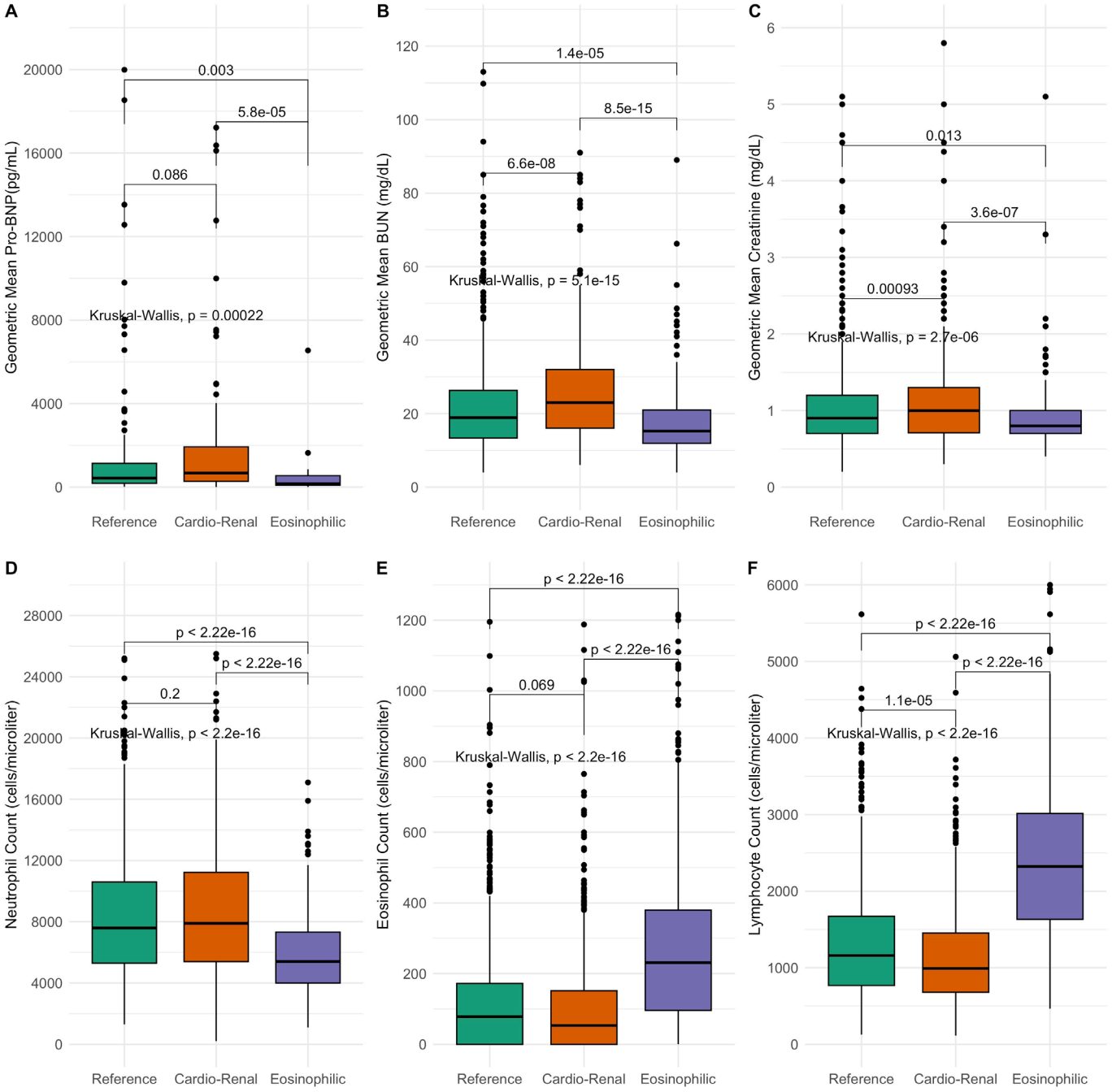
COPD=chronic obstructive pulmonary disease; BMI=body mass index

The high mortality rates of the cardio-renal subgroup led us to determine the survival times stratified by subgroups for severe AECOPD following hospitalization. This analysis showed that in contrast to the cardio-renal subgroup, the eosinophilic subgroup had the best median survival times after hospital discharge (Figures 2A and 2B).

To understand the relationship between COPD subgroups and the Rome criteria for severe AECOPDs,[21] we identified patients with respiratory acidosis based on arterial blood gas (ABG) testing (pH<7.35 and partial pressure of carbon dioxide>45mm Hg) at any point of their admission (n=65). There were no differences in severe AECOPDs across subgroups (Table 2).

To understand the factors that impact survival time in the COPD exacerbation subgroups, we first performed a univariate Cox regression analysis using subgroup, age, sex,

Figure 1. Cardiac, Renal Function, and Leukocyte Counts in COPD Subgroups



A. pro-BNP; B. BUN; C. Creatinine; D. Absolute neutrophil count; E. Absolute eosinophil count; F. Absolute lymphocyte count.

COPD=chronic obstructive pulmonary disease; pro-BNP=probrain natriuretic peptide; BUN=blood urea nitrogen

admission to the intensive care unit (ICU), heart failure, and chronic kidney disease given their potential influence on the subgroups and relevant biological input of age and sex. We found that subgroup assignment, age, ICU admission, and heart failure predicted survival time in the univariate analysis (Table 3). Because the hazard ratio distribution of absolute eosinophil counts crossed 1 in the univariate analysis, absolute eosinophil counts were not considered in the multivariate model. The multivariate Cox regression analysis included subgroup, age, admission to the ICU, and heart failure (Figure 2C and Table 3). After controlling for age, admission to the ICU, and heart failure, subgroup categories had a significant impact on survival.

A COVID-19 Cohort of COPD Patients Replicates the Original Subgroups

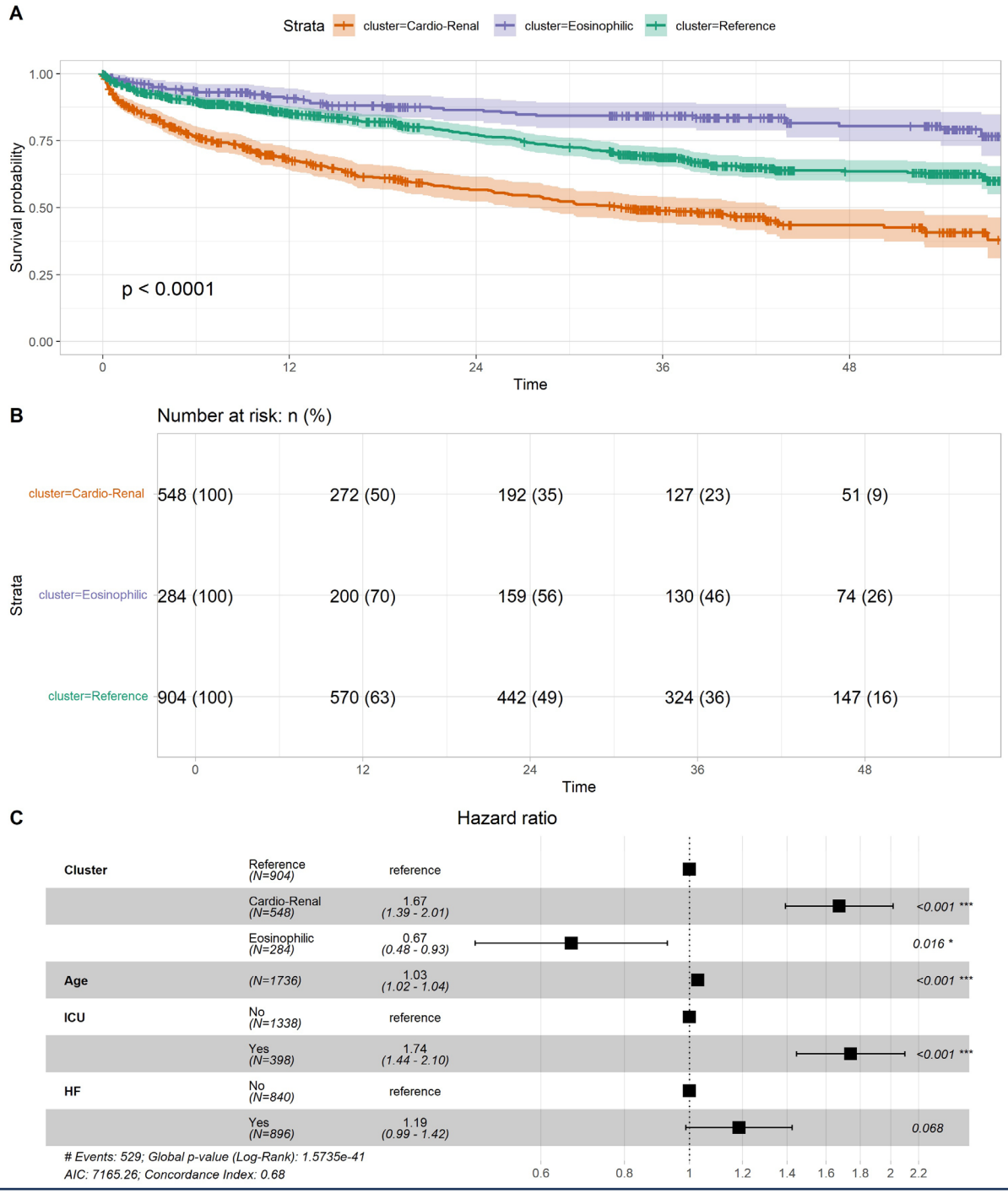
The triggers for severe AECOPDs that require hospitalization are heterogeneous, and their influence on the clustering of COPD exacerbations is unclear. SARS-CoV-2 infection, the causal agent of COVID-19, is an exceptional trigger for COPD exacerbations and disproportionately affects patients

Table 2. Outcomes of COPD Subgroups

	Reference (n=904)	Cardio-Renal (n=548)	Eosinophilic (n=284)	*P*-value
Length of Stay (days)	6.35 (3–7)	7.56 (3–9)	5.98 (2–6)	<0.01
Severe by Rome Criteria n (%)	33 (4)	25 (5)	7 (3)	0.3
Intensive Care Unit Admission n (%)	206 (23)	139 (25)	53 (19)	0.09
Death During Hospitalization n (%)	10 (1)	29 (5)	4 (1)	<0.01
Readmission Within 30 Days n (%)	57 (6)	27 (5)	22 (8)	0.26
Death in a Year n (%)	126 (14)	164 (30)	24 (9)	<0.01

COPD=chronic obstructive pulmonary disease; BMI=body mass index

Figure 2. Kaplan-Meier Curve and Cox-Proportional Hazards of COPD Subgroup Mortality



A. Kaplan-Meier curve for mortality by subgroup; B. Strata values for Kaplan-Meier curve; C. Cox-proportional hazards model for mortality.

ICU=intensive care unit; HF=heart failure; AIC=Akaike information criterion

Table 3. Univariate and Multivariate Analyses of Factors Associated With Mortality Risk

	Univariate Analysis		Multivariate Analysis	
	HR (95% CI)	*P*-value	HR (95% CI)	*P*-value
Subgroup				
Cardio-Renal	2.02 (1.70–2.43)	<0.01	1.67 (1.39–2.01)	<0.01
Eosinophilic	0.50 (0.36–0.69)	<0.01	0.67 (0.48–0.93)	0.02
Age	1.04 (1.03–1.05)	<0.01	1.03 (1.02–1.03)	<0.01
Intensive Care Unit Admission	1.70 (1.41–2.04)	<0.01	1.74 (1.44–2.10)	<0.01
Chronic Kidney Disease	1.13 (0.96–1.36)	0.13		
Heart Failure	1.58 (1.32–1.89)	<0.01	1.19 (0.99–1.42)	0.07
Absolute Eosinophil Count	0.99 (0.99–1)	0.04		

HR=hazard ratio; CI=constant interval

with COPD.[22] As a test of the validity of the severe COPD exacerbation subgroups, we implemented a deep learning classifier in a separate cohort of COPD patients in our hospital system admitted with COVID-19.

The 3 original AECOPD subgroups were recapitulated in this COVID-19 AECOPD cohort (n=1646) (Table 4). In the COVID-19 cohort, 68% of the patients were included in the reference subgroup, while 4% were classified as eosinophilic. There were no differences in sex or monocyte counts between subgroups in the COVID-19 cohort, similar to the original cohort. The cardio-renal subgroup in the COVID-19 cohort was the oldest (77 years [68–84]) and had the lowest BMI (27.6kg/m^2 [23.2–32.4]). Similarly to the cardio-renal subgroup in the original cohort, the COVID-19 cardio-renal subgroup had the highest prevalence of heart failure (60%) and chronic kidney disease (48%) of all 3 subgroups. The rates of antibiotic administration (75%) and systemic steroids (55%) were highest in this subgroup, in contrast to the original cohort (Table 1). Like the original cohort, the COVID-19 cardio-renal subgroup had the highest serum levels of pro-BNP, BUN, and creatinine (Figures 3A-C). Except for systemic steroids, used in the classifier to identify subgroups, no differences in tocilizumab or remdesivir use were seen across the COVID-19 subgroups (Table 4).

Remarkably, leukocyte counts in the COVID subgroups, also recapitulated the pattern seen in the original cohort, with the highest lymphocyte counts (1760 cells/microliter [1520–2260]) and eosinophil counts (127 cells/microliter [50–203]) in the eosinophilic subgroup. While the cardio-renal subgroup had elevated neutrophil counts (5380 cells/microliter [3591–7595]) and the lowest lymphocyte counts (900 cells/microliter [638–1203]).

Inflammatory Profiles of COVID-19 COPD Subgroups

To determine whether blood leukocyte counts seen in the COVID-19 subgroups were associated with distinct cytokine or inflammatory profiles, we compared the levels of 11 cytokines in the 3 subgroups. Following FDR adjustment, we found that 3 cytokines, interleukin (IL)-1beta, IL-2R, and IL-8, were differentially expressed (Table 4). The eosinophilic subgroup had the highest mean IL-1 beta values (Table 4), in keeping with previous studies describing IL-1beta release by eosinophils[23]; in contrast, the levels of IL-2R were lowest in the eosinophilic subgroup (Figure 4A).

Higher levels of IL-8, a cytokine involved in neutrophil recruitment and activation,[24] were associated with higher neutrophil counts in the reference and cardio-renal subgroups, compared to the eosinophilic subgroup (Figure 4B). Serum levels of the type 2 (T2) cytokines, IL-4, IL-5, and IL-13 were similar in the 3 subgroups (Supplementary Table 1 in the online supplement). Furthermore, serum levels of C-reactive protein (CRP) mirrored IL-2R, IL-8, and neutrophil counts in the 3 subgroups (Figure 4C). CRP levels ≥10mg/L which were included in the Rome proposal,[21] were more common in the reference and cardio-renal subgroups compared to the eosinophilic subgroup (Table 4). This suggests higher levels of inflammation in the COVID-19 reference and cardio-renal subgroups compared to the eosinophilic subgroup.

The Cardio-Renal Subgroup of the COVID-19 Cohort was Characterized by High Mortality

To determine whether associations between outcomes and subgroups were present in the COVID-19 cohort, we examined differences in ICU admission, severe AECOPD by Rome criteria based on their first ABG, 30-day readmission, length of stay, and hospital mortality between COPD subgroups. The rates of admission to the ICU and 30-day readmission were similar to those of the original subgroups (Table 5). Like the original subgroups, we found a shorter length of stay for the eosinophilic subgroup (6.9 days [4.1–12.1]). Although we lacked information beyond the hospitalization for COVID-19, the cardio-renal subgroup showed higher rates of inpatient mortality (26%), comparable to those in the cardio-renal subgroup of the original cohort (30%) within the first year after hospitalization.

Table 4. COPD Subgroups With COVID-19

	Reference (n=1114) 68%	Cardio-Renal (n=471) 29%	Eosinophilic (n=61) 4%	*P*-value
Age (years)	72 (61–81)	77 (68–84)	62 (52–65)	<0.01
Female Sex n (%)	605 (54.3)	266 (56.5)	42 (68.9)	0.07
BMI (kg/m^2)	29.1 (24.30–35.11)	27.6 (23.19–32.40)	30.4 (26.00–35.56)	<0.01
Comorbidities				
Heart Failure n (%)	503 (45)	281 (60)	26 (43)	<0.01
Cerebrovascular Disease n (%)	310 (28)	142 (30)	13 (21)	0.30
Diabetes Mellitus n (%)	642 (58)	250 (53)	34 (56)	0.25
Chronic Kidney Disease n (%)	408 (37)	224 (48)	20 (33)	<0.01
Lung Cancer n (%)	48 (4)	21 (5)	1 (2)	0.58
Hypertension n (%)	993 (89)	432 (92)	49 (80)	0.02
Multiple Comorbidities n (%)	872 (78)	392 (83)	45 (74)	0.12
Medications				
Albuterol n (%)	110 (10)	47 (10)	1 (2)	0.10
Antibiotic n (%)	710 (64)	354 (75)	29 (48)	<0.01
Inhaled Corticosteroids n (%)	12 (1)	6 (1)	0 (0)	0.66
Long-Acting Muscarinic Antagonist n (%)	179 (16)	114 (24)	15 (25)	<0.01
Leukotriene Receptor Antagonist n (%)	139 (13)	54 (12)	6 (10)	0.73
Nicotine Replacement n (%)	3 (0.27)	0 (0)	0 (0)	0.49
Systemic Steroids n (%)	557 (50)	259 (55)	16 (26)	<0.01
Ipratropium n (%)	85 (8)	39 (8)	1 (2)	0.18
Tocilizumab n (%)	225 (20)	104 (22)	13 (21)	0.69
Remdesivir n (%)	19 (2)	8 (2)	0	0.59
Complete Blood Count				
White Blood Cells (10^3 cells/μL)	6.98 (5.15–9.80)	7.21 (5.20–9.47)	6.75 (5.11–8.59)	0.51
Absolute Neutrophil Count (cells/μL)	5040 (3472–7685)	5380 (3591–7595)	3940 (2637–5182)	<0.01
Absolute Eosinophil Count (cells/μL)	33 (0–100)	29 (0–83)	127 (50–203)	<0.01
Absolute Basophil Count (cells/μL)	0 (0–17.6)	0 (0–16.7)	18.2 (0–66.7)	<0.01
Absolute Monocyte Count (cells/μL)	540 (400–721)	525 (381–692)	561 (400–782)	0.15
Absolute Lymphocyte Count (cells/μL)	999 (700–1420)	900 (638–1203)	1760 (1520–2260)	<0.01
Hematocrit (%)	37.4 (32.9–41.0)	34.8 (30.3–38.7)	36.3 (34.5–39.5)	<0.01
Hemoglobin (g/dL)	12.0 (10.5–13.2)	11.0 (9.39–12.4)	11.7 (10.6–12.7)	<0.01
Platelets (10^3 cells/μL)	218 (167–285)	212 (153–260)	261 (187–363)	<0.01
Inflammatory Markers				
IL-1 Beta (pg/mL)	1 (1–5) n=642	1 (1–5) n=278	5 (1–5) n=34	0.048[a]
IL-2R (pg/mL)	2288 (1357–3537) n=649	2390 (1394–4300) n=278	1562 (1057–2440) n=34	0.048[a]
IL-8 (pg/mL)	22.7 (5–44.3) n=644	23.4 (5–45.7) n=278	5 (5–25.2) n=34	0.048[a]
C-Reactive Protein (mg/dL)	6.3 (2.5–10.3) n=381	5.9 (2.9–11.3) n=158	2.1 (0.9–6.0) n=16	0.02
C-Reactive Protein Rome[b] n (%)	324 (29)	154 (33)	15 (25)	<0.01

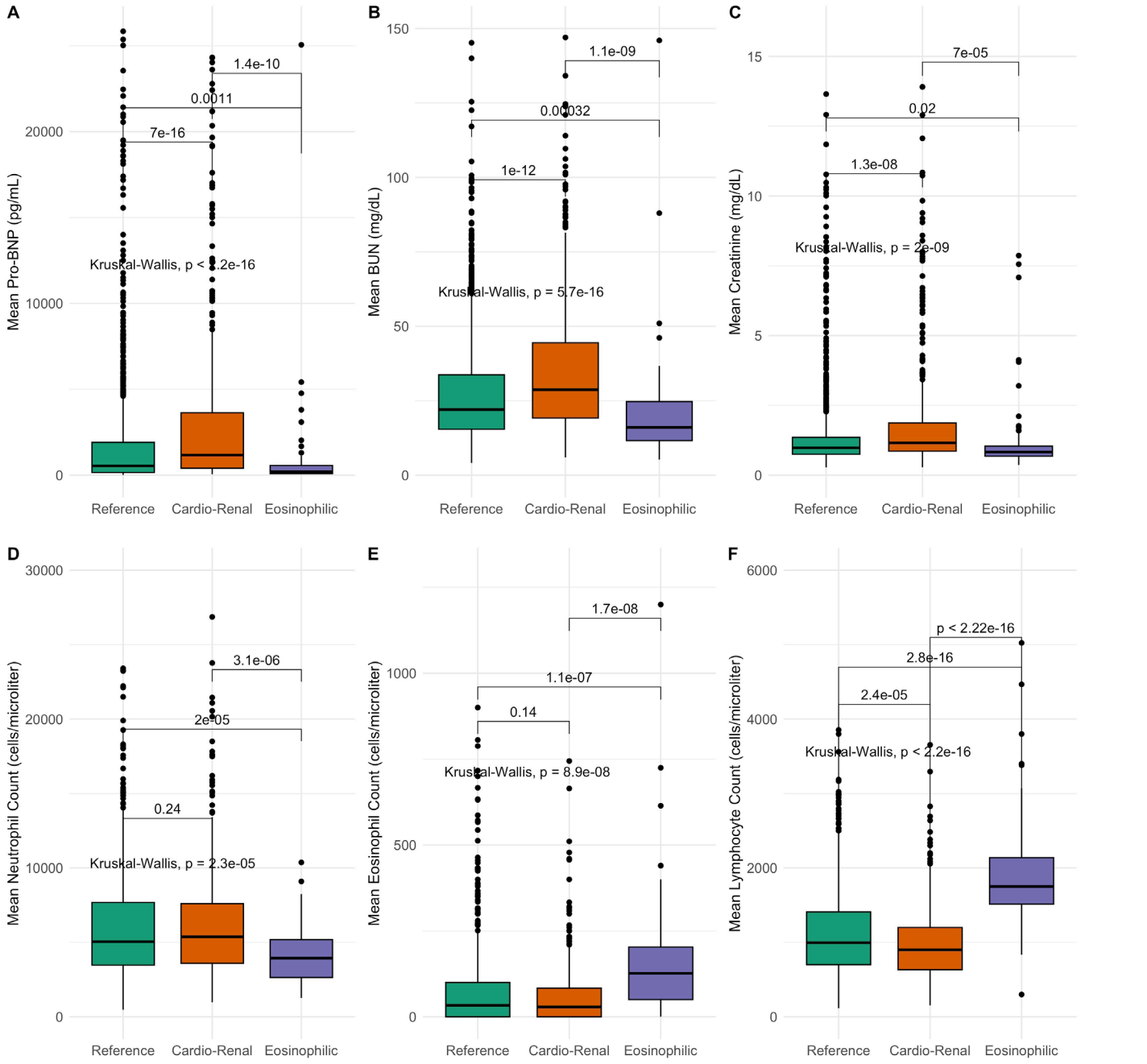
[a]FDR adjusted for 11 cytokines (IL-1 beta, IL-2, IL-2R, IL-4, IL-5, IL-6, IL-8, IL-10, IL-12, IL-13, IL-17, all values in Supplemental Table 1 in the online supplement)
[b]CRP ≥10mg/L.

COPD=chronic obstructive pulmonary disease; BMI=body mass index; IL=interleukin; FDR=false discovery rate; CRP=C-reactive protein

Discussion

We found 3 subgroups of severe AECOPDs using ML on EHR data from 3382 hospitalized patients. A total of 2 of the 3 subgroups were characterized by specific comorbidities or leukocyte profiles. First, a cardio-renal subgroup was associated with increased mortality during and after hospitalization for AECOPD. This was followed by an eosinophilic subgroup that had the shortest hospital stay, suggesting a milder pattern of exacerbation. It is notable that the subgroups were evident despite differences between the cohorts, including triggers for hospitalization. In the original cohort, the triggers were not captured by our study design, while the second cohort was restricted to patients hospitalized with COVID-19. Overall, these findings demonstrate that these subgroups are stable and support the use of ML classifiers in EHRs to classify hospitalizations with AECOPDs. Increasing automated recognition of AECOPD subphenotypes in EHRs presents a clinical opportunity to develop precision medicine interventions to improve disease outcomes.

Figure 3. Validation of COPD Subgroups in Patients Admitted for COVID-19 at Yale-New Haven Hospital System



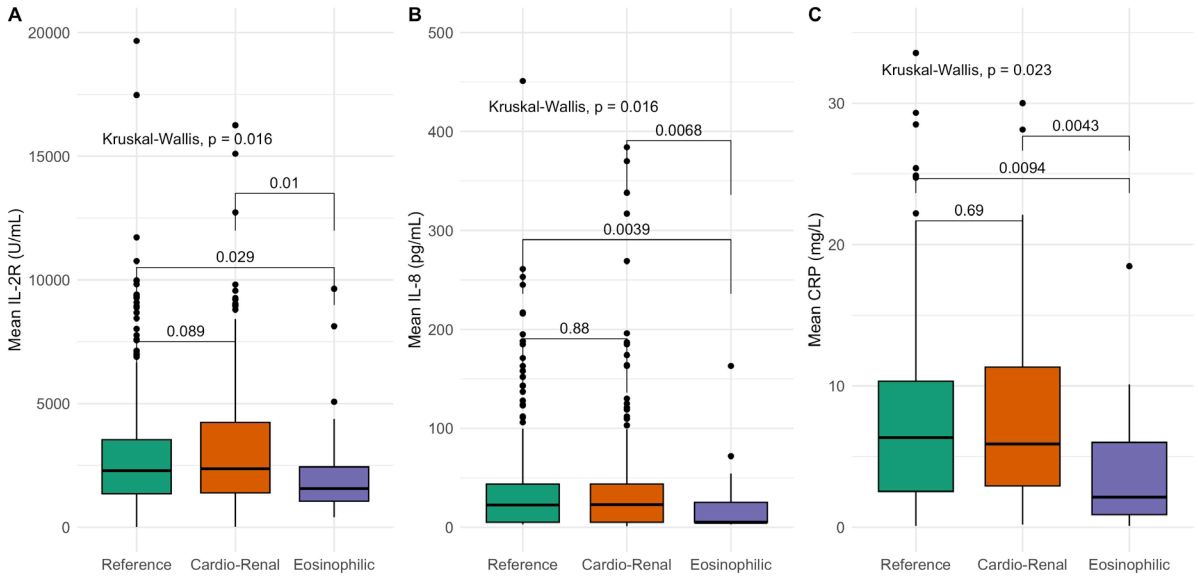
A. pro-BNP; B. BUN; C. Creatinine; D. Absolute neutrophil count; E. Absolute eosinophil count; F. Absolute lymphocyte count.

COPD=chronic obstructive pulmonary disease; pro-BNP=probrain natriuretic peptide; BUN=blood urea nitrogen

These subgroups are important for their morbidity and mortality, as well as their specific clinical characteristics. The cardio-renal subgroup not only recapitulates what is known about the impact of specific comorbidities on COPD outcomes,[8] it also captures other phenotypic traits associated with increased mortality, including a lower BMI.[25] The identification of lower lymphocyte counts combined with higher neutrophil counts in this subgroup is also consistent with multiple studies that examined the neutrophil to lymphocyte ratio in AECOPDs as a marker of exacerbation risk and mortality.[26] Considering the aging process, the presence of COPD, chronic cardiac and renal disease, and the presence of unique inflammation surrogates in neutrophils and lymphocytes, it is plausible that mechanisms of immunosenescence may be present in this subgroup.[27] Recapitulating all these features associated with poor outcomes into a single subgroup strengthens our ability to understand this phenotype and can aid in the identification of AECOPD triggers and therapeutic targets unique to this group of patients.

We identified the eosinophilic subgroup in the

Figure 4. Cytokine and Inflammatory Profiles of COPD Subgroups With COVID-19



A. IL-2R; B. IL-8; C. CRP

COPD=chronic obstructive pulmonary disease; IL=interleukin; CRP=c=reactive protein

original cohort through the integration of comorbidities associated with T2 inflammation and blood counts. Despite the confounding effect of systemic steroid administration on blood eosinophil counts, the ability to identify this subgroup points to the robustness of blood eosinophils as a marker to distinguish this subgroup. This subgroup was also characterized by milder exacerbations characterized by shorter length of stay, consistent with previous studies of AECOPDs requiring hospitalization.[17] These differences are likely related to age, among other factors. We speculate that it is possible that this subgroup of exacerbations is more responsive to the administration of systemic steroids. We did not see differences in T2 cytokines in the validation cohort, and this may reflect limited power to identify differences or the influence of concomitant viral infection and COVID therapies. Furthermore, the demonstration of clinical benefit in COPD with increased blood eosinophils after dual blockade of the IL4/IL13 T2 pathway with dupilumab[11] confirms this as a distinct endotype based both on molecular mechanism and response to treatment.[4]

The largest reference cluster had a mix of clinical features and outcomes that fell between the cardio-renal and eosinophilic subgroups. This suggests that there are additional AECOPD phenotypes that are not captured by the current parameters of our analysis. For instance, key differences in the diagnosis of heart failure, including ejection fraction and the mechanisms involved including diastolic and systolic failure, are essential for more accurate classification. Our study was intended as a proof-of-concept for computable subgroups of severe AECOPD, which led to the use of conservative clustering parameters to prevent overclustering of subgroups, which may lead to the identification of very small groups without broad applicability. The results of future studies may identify new subgroups using different parameters.

We recognize the limitations of our model. These include the lack of spirometric values to define COPD, background therapies, and lung imaging patterns in which subgroups were defined. The single hospital system and selected EHR features may contribute to selection bias. The differences between subgroups may also have been driven by specific molecular determinants that EHRs failed to capture. To address some of these limitations, we used strict criteria to define COPD including multiple International Classification of Diseases-Tenth Revision entries, excluding those with dual diagnoses of asthma and COPD, and use of complete, routinely available clinical data rather than imputed values. To make a similar model applicable to other centers, we carefully selected data on inpatient medication administration profiles and structured data when available. Finally, our dataset did not collect all the variables required by the Rome proposal to determine degrees of severity of AECOPDs. We sought to overcome this limitation by focusing on the severe category defined by ABG testing in a subset of patients. It is expected that subsequent iterations of our current approach will refine the role of computable subgroups in COPD classification.

Conclusions

Computable subphenotypes of severe AECOPD identify a cardio-renal subgroup associated with increased mortality. This subgroup includes several known features connected to poor outcomes in COPD. In contrast, a separate eosinophilic

Table 5. COVID-19 COPD Subgroup Outcomes

	Reference (n=1114)	Cardio-Renal (n=471)	Eosinophilic (n=61)	*P*-value
Length of Stay (days)	7.9 (4.9–14.0)	8.8 (4.9–15.3)	6.9 (4.1–12.1)	0.03
Severe by Rome Criteria n (%)	81 (7)	44 (9)	5 (8)	0.43
Intensive Care Unit Admission n (%)	229 (20.6)	110 (23.4)	9 (14.8)	0.21
Readmission Within 30 Days n (%)	155 (13.9)	54 (11.5)	6 (9.8)	0.31
Death During Hospitalization n (%)	180 (16.2)	122 (25.9)	2 (3.3)	<0.01

COPD=chronic obstructive pulmonary disease

subgroup is associated with milder AECOPD requiring hospitalization. ML can be used to improve patient classification using data collected on EHRs and result in new treatment paradigms tailored to specific disease subtypes.

Acknowledgments

Author contributions: HL and JLG contributed to the conception and design, data acquisition, and analysis. All authors contributed to the final article drafting and revision and gave final approval.

Other acknowledgments: The author wish to acknowledge the assistance and expertise of Richard Hintz and Krishna Daggula at Yale's Joint Data Analytics Team.

Declaration of Interests

HL and JH have no conflicts of interest related to this work. JZ reports funding from a National Institutes of Health (NIH) training grant (2T32HL007778-26), and personal fees for participation in practice update. SK reports funding from an NIH training grant (2T32HL007778-26). MS reports funding from the NIH/National Heart, Lung, and Blood Institute (NHLBI) (R01 HL155948). JG reports funding from the NIH/NHLBI (R01 HL153604 and R03 HL154275).

References

1. Castaldi PJ, Dy J, Ross J, et al. Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. *Thorax*. 2014;69(5):416-423. https://doi.org/10.1136/thoraxjnl-2013-203601

2. Ghebre MA, Bafadhel M, Desai D, et al. Biological clustering supports both "Dutch" and "British" hypotheses of asthma and chronic obstructive pulmonary disease. *J Allergy Clin Immunol*. 2015;135(1):63-72. https://doi.org/10.1016/j.jaci.2014.06.035

3. Sakornsakolpat P, Prokopenko D, Lamontagne M, et al. Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nat Genet*. 2019;51:494-505. https://doi.org/10.1038/s41588-018-0342-2

4. Anderson GP. Endotyping asthma: new insights into key pathogenic mechanisms in a complex, heterogeneous disease. *Lancet*. 2008;372(9643):1107-1119. https://doi.org/10.1016/S0140-6736(08)61452-X

5. Lopez-Campos JL, Agustí A. Heterogeneity of chronic obstructive pulmonary disease exacerbations: a two-axes classification proposal. *Lancet Respir Med*. 2015;3(9):729-734. https://doi.org/10.1016/S2213-2600(15)00242-8

6. Perera PN, Armstrong EP, Sherrill DL, Skrepnek GH. Acute exacerbations of COPD in the United States: inpatient burden and predictors of costs and mortality. *COPD*. 2012;9(2):131-141. https://doi.org/10.3109/15412555.2011.650239

7. Shah T, Churpek MM, Coca Perraillon M, Konetzka RT. Understanding why patients with COPD get readmitted: a large national study to delineate the Medicare population for the readmissions penalty expansion. *Chest*. 2015;147(5):1219-1226. https://doi.org/10.1378/chest.14-2181

8. Vanfleteren LEGW, Spruit MA, Groenen M, et al. Clusters of comorbidities based on validated objective measurements and systemic inflammation in patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2013;187(7):728-735. https://doi.org/10.1164/rccm.201209-1665OC

9. Haghighi B, Choi S, Choi J, et al. Imaging-based clusters in current smokers of the COPD cohort associate with clinical characteristics: the SubPopulations and Intermediate Outcome Measures in COPD Study (SPIROMICS). *Respir Res*. 2018;19:178. https://doi.org/10.1186/s12931-018-0888-7

10. Pavord ID, Chanez P, Criner GJ, et al. Mepolizumab for eosinophilic chronic obstructive pulmonary disease. *N Engl J Med*. 2017;377(17):1613-1629. https://doi.org/10.1056/NEJMoa1708208

11. Bhatt SP, Rabe KF, Hanania NA, et al. Dupilumab for COPD with type 2 inflammation indicated by eosinophil counts. *N Engl J Med*. 2023;389(3):205-214. https://doi.org/10.1056/NEJMoa2303951

12. Henry J, Pylypchuk Y, Searcy T, Patel V. Adoption of electronic health record systems among US non-federal acute care hospitals: 2008-2015. HealthIT website. Published 2016. Accessed 2024. https://www.healthit.gov/sites/default/files/briefs/2015_hospital_adoption_db_v17.pdf

13. EMC Digital Universe, IDC. Vertical industry brief: digital universe driving data growth in healthcare. Cyclone Interactive website. Published 2014. Accessed 2024. https://www.cycloneinteractive.com/sites/cyclone/assets/File/digital-universe-healthcare-vertical-report-ar.pdf

14. Richesson RL, Wiley LK, Gold S, et al. Electronic health records-based phenotyping. Introduction. Rethinking Clinical Trials website. Published 2020. Accessed March 2024. https://rethinkingclinicaltrials.org/chapters/conduct/electronic-health-records-based-phenotyping/electronic-health-records-based-phenotyping-introduction/

15. Lopez K, Li H, Lipkin-Moore Z, et al. Deep learning prediction of hospital readmissions for asthma and COPD. *Respir Res*. 2023;24:311. https://doi.org/10.1186/s12931-023-02628-7

16. Yale University, Department of Medicine. COVID-19 Explorer and Repository tool: DOM-CovX. Yale University website. Published 2020-2021. Accessed June 27, 2021. https://spinup-0011f4.spinup.yale.edu/domcovx/

17. Bafadhel M, Greening NJ, Harvey-Dunstan TC, et al. Blood eosinophils and outcomes in severe hospitalized exacerbations of COPD. *Chest*. 2016;150(2):320-328. https://doi.org/10.1016/j.chest.2016.01.026

18. Vedel-Krogh S, Nielsen SF, Lange P, Vestbo J, Nordestgaard BG. Blood eosinophils and exacerbations in chronic obstructive pulmonary disease. The Copenhagen General Population Study. *Am J Respir Crit Care Med*. 2016;193(9):965-974. https://doi.org/10.1164/rccm.201509-1869OC

19. Yun JH, Lamb A, Chase R, et al. Blood eosinophil count thresholds and exacerbations in patients with chronic obstructive pulmonary disease. *J Allergy Clin Immunol*. 2018;141(6):2037-2047.e10. https://doi.org/10.1016/j.jaci.2018.04.010

20. Roberts CM, Stone RA, Lowe D, Pursey NA, Buckingham RJ. Co-morbidities and 90-day outcomes in hospitalized COPD exacerbations. *COPD*. 2011;8(5):354-361. https://doi.org/10.3109/15412555.2011.600362

21. Celli BR, Fabbri LM, Aaron SD, et al. An updated definition and severity classification of chronic obstructive pulmonary disease exacerbations: the Rome proposal. *Am J Respir Crit Care Med*. 2021;204(11):1251-1258. https://doi.org/10.1164/rccm.202108-1819PP

22. Gerayeli FV, Milne S, Cheung C, et al. COPD and the risk of poor outcomes in COVID-19: a systematic review and meta-analysis. *EClinicalMedicine*. 2021;33:100789. https://doi.org/10.1016/j.eclinm.2021.100789

23. Esnault S, Kelly EAB, Nettenstrom LM, Cook EB, Seroogy CM, Jarjour NN. Human eosinophils release IL-1ß and increase expression of IL-17A in activated CD4+ T lymphocytes. *Clin Exp Allergy*. 2012;42(12):1756-1764. https://doi.org/10.1111/j.1365-2222.2012.04060.x

24. Rajarathnam K, Sykes BD, Kay CM, et al. Neutrophil activation by monomeric interleukin-8. *Science*. 1994;264(5155):90-92. https://doi.org/10.1126/science.8140420

25. Hallin R, Gudmundsson G, Suppli Ulrik C, et al. Nutritional status and long-term mortality in hospitalised patients with chronic obstructive pulmonary disease (COPD). *Respir Med*. 2007;101(9):1954-1960. https://doi.org/10.1016/j.rmed.2007.04.009

26. Paliogiannis P, Fois AG, Sotgia S, et al. Neutrophil to lymphocyte ratio and clinical outcomes in COPD: recent evidence and future perspectives. *Eur Respir Rev*. 2018;27(147):170113. https://doi.org/10.1183/16000617.0113-2017

27. Murray MA, Chotirmall SH. The impact of immunosenescence on pulmonary disease. *Mediators Inflamm*. 2015;692546. https://doi.org/10.1155/2015/692546