Original Research

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

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

[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

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

Running Head: **Phenomapping Severe COPD Exacerbations**

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

Abbreviations: COPD: Chronic obstructive pulmonary disease; AECOPD: Acute exacerbations of COPD; ML: Machine learning; EHR: Electronic health records; YNHHS: Yale-New Haven Hospital System; FDR: False discovery rate; ICS/LABA: Inhaled corticosteroid/Long-acting beta-agonist combination; BMI: Body mass index (BMI); pro-BNP: pro-brain natriuretic peptide(); BUN: Blood urea nitrogen; 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 NIH.

Date of Acceptance: October 13, 2024 | ***Publication Online Date***: October 17, 2024

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; Published online October 17, 2024.
https://doi.org/10.15326/jcopdf.2024.0556

This article has an online data supplement.

Abstract

Rationale: Acute exacerbations of COPD (AECOPD) 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: Determine whether ML implementation using EHR data from severe AECOPD requiring hospitalization identifies relevant subgroups.

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

Measurements and main results: We identified three subgroups in the non-COVID cohort (n=1,736). 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 IL-1 beta, IL-2R, and IL-8 levels (FDR ≤ 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.

Published online October 17, 2024 https://doi.org/10.15326/jcopdf.2024.0556

Introduction

Chronic obstructive pulmonary disease (COPD) is heterogeneous[1–3]. Factors involved in heterogeneity include clinical characteristics, distinct pathobiological characteristics, including types of inflammation and 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 (AECOPD) also exhibit this heterogeneity, which can be related to the baseline characteristics of the subgroups or to the triggers for exacerbations [5].

Severe AECOPD that requires hospitalization are associated with significant morbidity and mortality, in addition to significant healthcare 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 AECOPD on patients and the healthcare system and underscore the importance of understanding the heterogeneity associated with severe exacerbations.

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

COPD subgroups have been identified by cytokine profiles [2]; a combination of clinical data including comorbidities [8]; a combination of clinical, physiologic, and imaging data[1]; and 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 healthcare providers to adopt electronic health records (EHRs). Currently, more than 95% of US hospitals have adopted EHRs [12]. As a result of EHR adoption the volume of healthcare data has increased exponentially from 153 Exabytes in 2013 to 2314 Exabytes in 2020 [13]. This massive increase in data encodes millions of healthcare 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 machine learning in hospitalizations for severe AECOPD 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 (YNHH) between September 30, 2012, after the Epic EHR system (Verona, WI) was implemented, and December 31, 2017. The Yale University Human Research

Protection Program approved this study. 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 (DOM-CovX) 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 (https://spinup-0011f4.spinup.yale.edu/domcovx/ (2021), Accessed 27th Jun 2021). 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 three. 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.

Copyright Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation ©2024
Published online October 17, 2024 https://doi.org/10.15326/jcopdf.2024.0556

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.

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.

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 1,736 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 three distinct subgroups in the resulting analysis (Table 1). Across all three subgroups, sex and absolute monocyte counts were similar, suggesting that sex or monocytes were not key factors in this classification.

Clinical characteristics of the AECOPD 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 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 three 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 three 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 three subgroups. This subgroup 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 AECOPD 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 pro-brain natriuretic peptide (pro-BNP), blood urea nitrogen (BUN), and creatinine values from patients in the three subgroups. The cardio-renal subgroup had the highest combined pro-BNP, BUN, and creatinine values of the three 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 (5,400 cells/microliter [4,000-7,300]), highest blood lymphocyte (2,325 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 [16–18], 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 [19], eosinophilic inflammation in COPD exacerbations [16] 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[16], we found that the eosinophilic subgroup had the shortest stay (5.98 days [2-6])(Table 2). During

Published online October 17, 2024 https://doi.org/10.15326/jcopdf.2024.0556

hospitalization (5%) and in the year following an AECOPD hospitalization (30%), the cardio-renal subgroup had the highest mortality rates.

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 AECOPD[20], we identified patients with respiratory acidosis based on arterial blood gas (ABG) testing (pH<7.35 and PaCO2 >45 mm Hg) at any point of their admission (n=65). There were no differences in severe AECOPD 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, 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 (Fig 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 AECOPD 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 affected patients with COPD [21]. 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 three original AECOPD subgroups were recapitulated in this COVID-19 AECOPD cohort (n=1,646) (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 (27.6 kg/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 three 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).

Published online October 17, 2024 https://doi.org/10.15326/jcopdf.2024.0556

Remarkably, leukocyte counts in the COVID subgroups, also recapitulated the pattern seen in the original cohort, with the highest lymphocyte counts (1,760 cells/microliter [1,520-2,260]) and eosinophil counts (127 cells/microliter [50-203)) in the eosinophilic subgroup. While the cardio-renal subgroup had elevated neutrophil counts (5,380 cells/microliter [3,591-7,595]) and the lowest lymphocyte counts (900 cells/microliter [638-1,203]).

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 three subgroups. Following FDR adjustment, we found that three cytokines, 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 [22]; 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 [23], 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 three subgroups (Supplementary Table 1). Furthermore, serum levels of C-reactive protein (CRP) mirrored IL-2R, IL-8, and neutrophil counts in the three subgroups (Figure 4C). CRP levels ≥ 10 mg/L which were included in the Rome proposal [20], 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 admission 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.

Discussion

We found three subgroups of severe AECOPD using machine learning on EHR data from 3,382 hospitalized patients. Two of the three 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 machine learning classifiers in EHRs to classify hospitalizations with AECOPD. Increasing automated recognition of AECOPD subphenotypes in EHRs presents a clinical opportunity to develop precision medicine interventions to improve disease outcomes.

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 [24]. 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 AECOPD as a marker of exacerbation risk and mortality [25]. 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 [26]. 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 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.

Published online October 17, 2024 https://doi.org/10.15326/jcopdf.2024.0556

This subgroup was also characterized by milder exacerbations characterized by shorter length of stay, consistent with previous studies of AECOPD requiring hospitalization [16]. 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, 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 ICD-10 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 AECOPD. 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 subgroup is associated with milder AECOPD requiring hospitalization. Machine learning can be used to improve patient classification using data collected on EHRs and result in new treatment paradigms tailored to specific disease subtypes.

Published online October 17, 2024 https://doi.org/10.15326/jcopdf.2024.0556

Acknowledgments

Richard Hintz and Krishna Daggula at JDAT

Declarations

Ethical approval was obtained from Yale University's Institutional Review Board (IRB).

Consent to participate

This project was approved by the IRB under a Waiver of Consent.

Consent for publication

Not applicable.

Author contributions

Conception and design H.L., J.L.G. Data acquisition, and analysis: H.L., J.L.G. Article drafting/revision: All authors Final approval: all authors.

Published online October 17, 2024 https://doi.org/10.15326/jcopdf.2024.0556

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):415-422.

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*. Published online August 2014. doi: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(3):494-505.

4. Anderson GP. Endotyping asthma: new insights into key pathogenic mechanisms in a complex, heterogeneous disease. *Lancet*. 2008;372(9643):1107-1119.

5. Lopez-Campos JL, Agustí A. Heterogeneity of chronic obstructive pulmonary disease exacerbations: a two-axes classification proposal. *The Lancet Respiratory Medicine*. 2015;3(9):729-734.

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.

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.

8. Vanfleteren LE, 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.

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(1):178.

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.

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.

12. Henry J, Pylypchuk Y, Searcy T, Patel V. Adoption of electronic health record systems among US non-federal acute care hospitals: 2008-2015. 2016. *URl: https://dashboard healthit gov/evaluations/databriefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015 php*. Published online 2018. https://dashboard.healthit.gov/evaluations/data-

briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php

13. Digital Universe Healthcare Report. IDC Healthcare. Accessed February 15, 2020. https://www.emc.com/analyst-report/digital-universe-healthcare-vertical-report-ar.pdf

14. Introduction. Rethinking Clinical Trials. June 30, 2020. Accessed March 7, 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(1):311.

16. 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. doi:10.1016/j.chest.2016.01.026

17. 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.

18. 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.

19. 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.

20. 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.

21. 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.

22. 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.

23. Rajarathnam K, Sykes BD, Kay CM, et al. Neutrophil activation by monomeric interleukin-8. *Science*. 1994;264(5155):90-92.

24. 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.

25. 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). doi:10.1183/16000617.0113-2017

26. Murray MA, Chotirmall SH. The Impact of Immunosenescence on Pulmonary Disease. *Mediators Inflamm*. 2015;2015:692546.

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
Body mass index (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

Published online October 17, 2024 https://doi.org/10.15326/jcopdf.2024.0556

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 beta-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

Published online October 17, 2024 https://doi.org/10.15326/jcopdf.2024.0556

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

Published online October 17, 2024 https://doi.org/10.15326/jcopdf.2024.0556

Table 3. Univariate and multivariate analyses of factors associated with mortality risk.

Variables	**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		

Published online October 17, 2024 https://doi.org/10.15326/jcopdf.2024.0556

Table 4. COPD subgroups with COVID-19

	Reference (n=1,114) 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
Body mass index (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 corticosteroid n (%)	12 (1)	6 (1)	0 (0)	0.66
Long-acting muscarinic antagonist n (%)	179 (16)	114 (24)	15 (25)	< 0.01

Published online October 17, 2024 https://doi.org/10.15326/jcopdf.2024.0556

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*
IL-2R (pg/mL)	2,288 (1,357-3,537) n=649	2,390 (1,394-4,300) n=278	1,562 (1,057-2,440) n=34	0.048*
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*
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** n (%)	324 (29)	154 (33)	15 (25)	<0.01

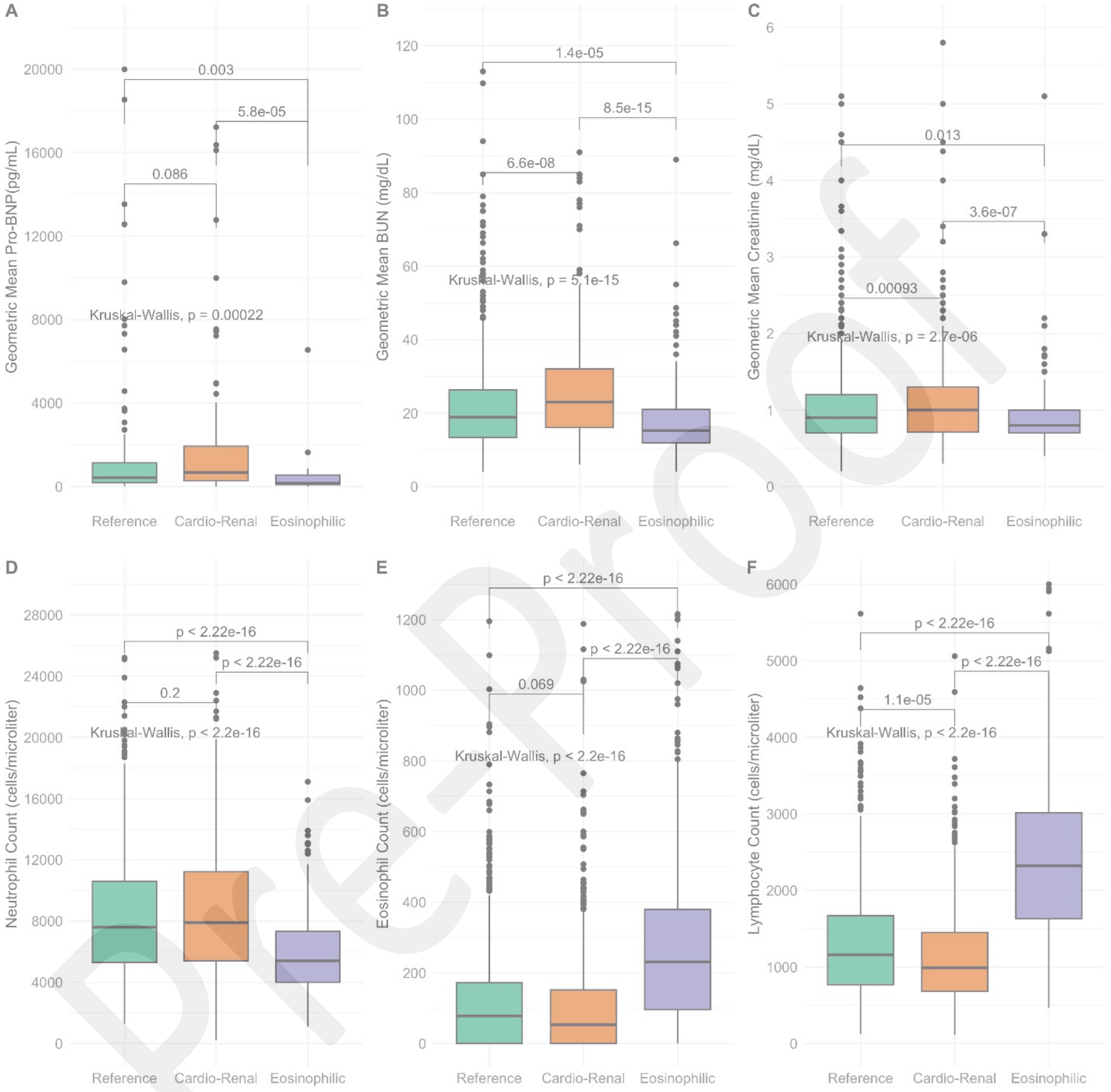
*FDR adjusted for 11 cytokines (IL-1 beta, IL-2, IL-2R, IL-4, IL-5, IL-6, IL-8, IL-10, IL-12, IL-13, IL-17, all values in supplemental Table 1) **CRP ≥ 10 mg/L.

Table 5. COVID-19 COPD subgroup outcomes

	Reference (n=1,114)	**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

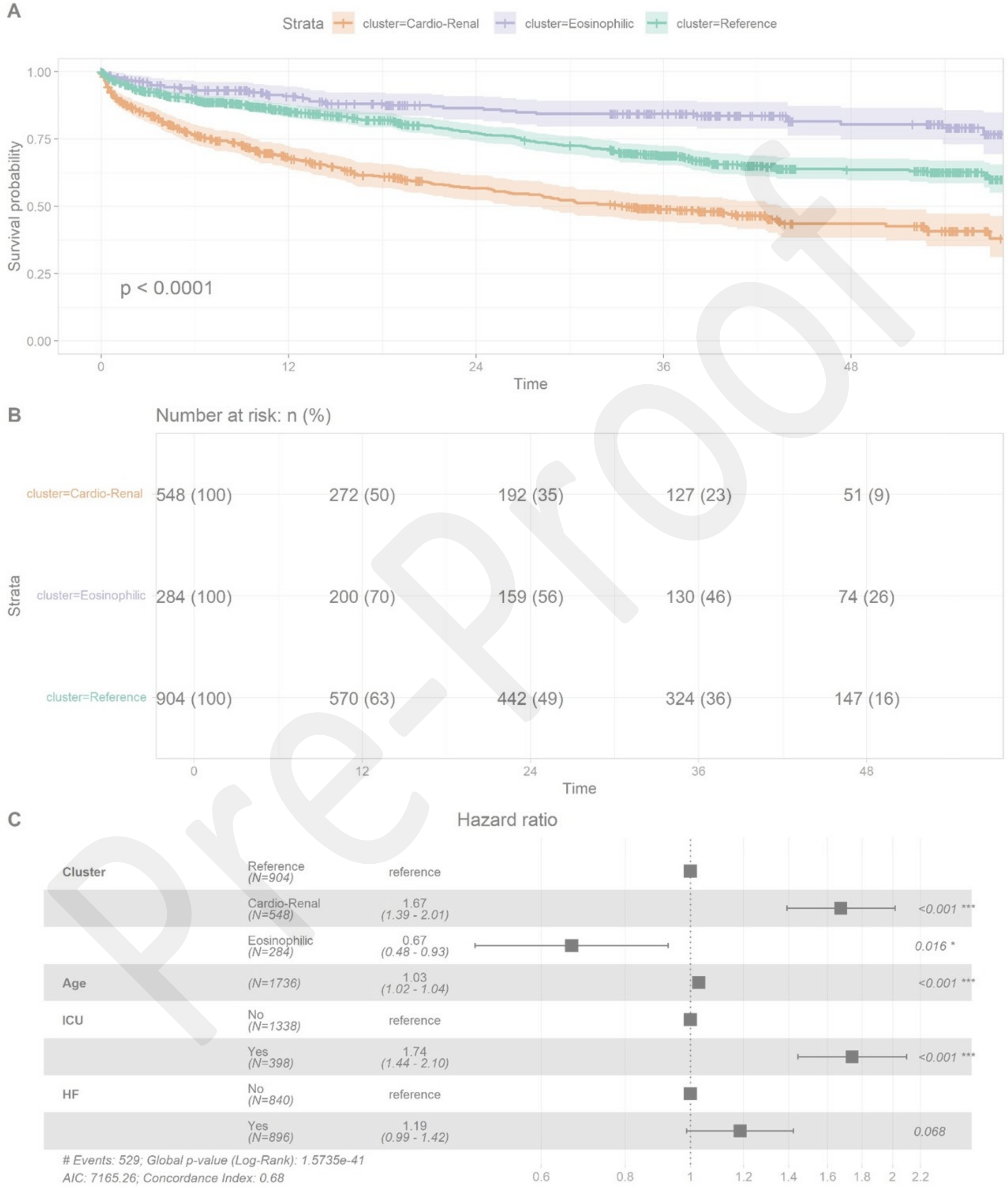
Published online October 17, 2024 https://doi.org/10.15326/jcopdf.2024.0556

Figure 1. Cardiac, renal function, and leukocyte counts in COPD Subgroups.
A. Pro-brain natriuretic peptide (pro-BNP). B. Blood urea nitrogen (BUN). C. Creatinine. D. Absolute neutrophil count. E. Absolute eosinophil count. F. Absolute lymphocyte count.



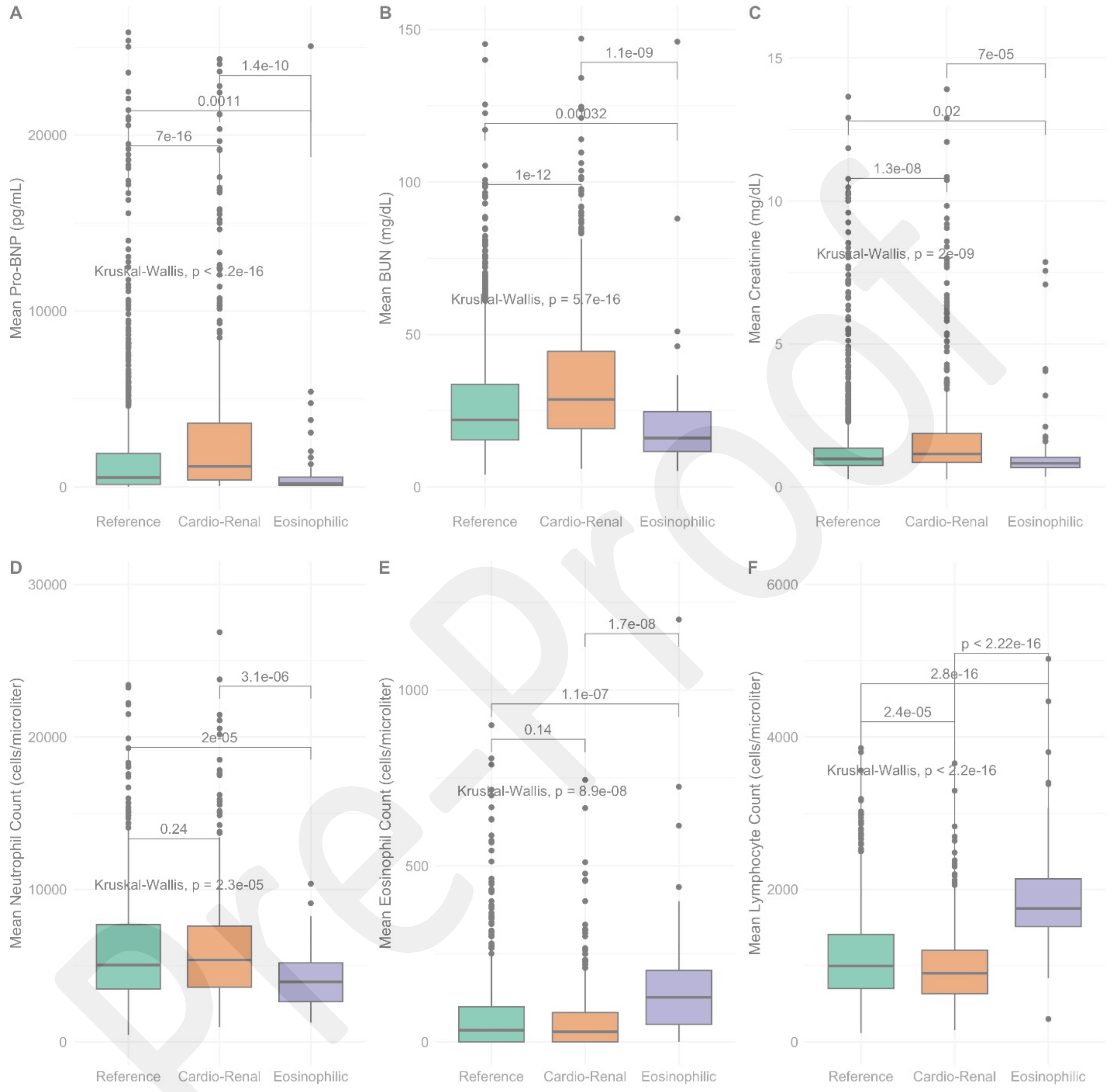
Published online October 17, 2024 https://doi.org/10.15326/jcopdf.2024.0556

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.



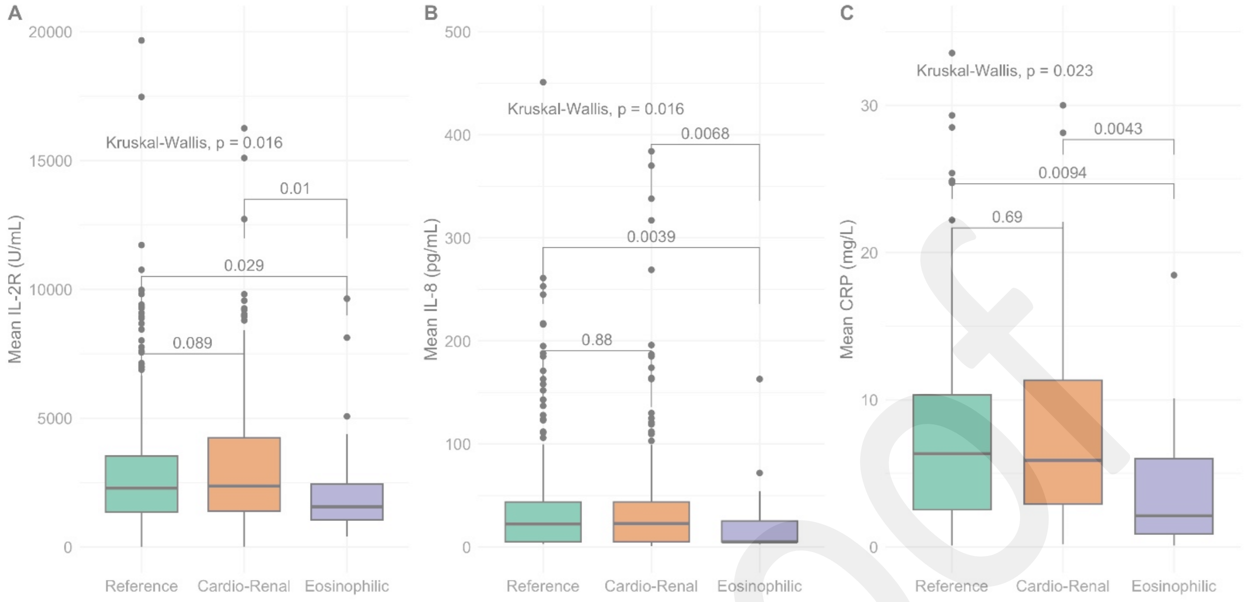
Published online October 17, 2024 https://doi.org/10.15326/jcopdf.2024.0556

Figure 3. Validation of COPD Subgroups in patients admitted for COVID-19 at YNHHS. A. Pro-brain natriuretic peptide (pro-BNP). B. Blood urea nitrogen (BUN). C. Creatinine. D. Absolute neutrophil count. E. Absolute eosinophil count. F. Absolute lymphocyte count.



Published online October 17, 2024 https://doi.org/10.15326/jcopdf.2024.0556

Figure 4. Cytokine and inflammatory profiles of COPD Subgroups with COVID-19.
A. IL-2R. B. IL-8. C. C-reactive protein (CRP).



Published online October 17, 2024 https://doi.org/10.15326/jcopdf.2024.0556

Online Supplement

Supplementary Methods

Original Cohort Data Source and Study Population

YNHH is a tertiary-care hospital with 1541 beds and two campuses in New Haven, Connecticut, USA. We included all participants who met the following criteria during the study period: This study was limited to hospital admissions from patients 18 years and older. The International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) codes ICD-10-CM code J44.1 and also patients with any combination of acute bronchitis (J20.90) with chronic obstructive pulmonary disease, unspecified (J44.9), unspecified chronic bronchitis (J42), and emphysema, unspecified (J43.9). Patients with dual ICD-10-CM diagnosis for asthma and COPD at the time of the index or subsequent hospitalizations were also excluded from these analyses. Data from the first hospitalization that met inclusion and exclusion criteria during the study period were used for this analysis. Following patient identification, demographic data, including self-reported race and self-reported Hispanic ethnicity, comorbidities, first measured laboratory data and inpatient medication records were collected from the index hospitalization. Comorbidity data were extracted from the EHR's past medical history section. Laboratory data included complete blood count and blood chemistries.

COVID Cohort

The Yale Department of Medicine COVID-19 Explorer and Repository tool (DOM-CovX) 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 (https://spinup-0011f4.spinup.yale.edu/domcovx/ (2021), Accessed 27th Jun 2021). These hospitals included YNHH, Greenwich Hospital (Greenwich, CT), Bridgeport Hospital (Bridgeport, CT), Lawrence & Memorial Hospital (New London, CT), and Westerly Hospital (Westerly, RI). The patients had a positive test for SARS-CoV-2 using reverse transcriptase–polymerase chain reaction assays performed on nasopharyngeal swab specimens. The samples were collected within 14 days after admission. Following patient identification, demographics, past medical history, self-reported race and ethnicity, smoking history (in pack years), admission vital signs, laboratory tests, medications administered during hospitalization, and discharge status (to assess in-hospital mortality), were retrieved

Clustering

The 51 training features included: Age, sex, body mass index (BMI), systolic blood pressure, diastolic blood pressure, mean arterial pressure, smoking status (former, current, never), insurance, coronary artery disease (CAD), cerebrovascular disease, diabetes mellitus (DM), chronic kidney disease (CKD), heart failure (HF), allergic rhinitis, nasal polyposis, gastroesophageal reflux (GERD), lung cancer, sleep apnea, and hypertension (HTN). Inpatient administration of albuterol, inhaled corticosteroids (ICS) with long-acting beta-agonist (LABA), long-acting muscarinic antagonist (LAMA), antibiotic, ipratropium, systemic steroid, ICS-only, leukotriene receptor antagonist (LTRA), theophylline, influenza vaccine, oseltamivir, pneumococcal vaccine, and nicotine supplementation. White blood cell (WBC) count, neutrophil percentage, eosinophil percentage, basophil percentage, monocyte percentage, lymphocyte percentage, absolute neutrophil count, absolute eosinophil count, absolute basophil count, absolute monocyte count, absolute lymphocyte count, red blood cell count (RBCC), hematocrit, hemoglobin, mean corpuscular hemoglobin (MHC), mean corpuscular hemoglobin concentration (MCHC), mean corpuscular volume (MCV), platelets, and mean platelet volume (MPV).

The next step involved categorizing the variables into numerical and string types.

Statistical Analysis

The R statistical software was used for statistical analyses. The packages dplyr v.0.8.5, lubridate v.1.7.9, survival v.3.5-3, survminer v.0.4.9, ggfortify v.0.4.12, ggplot2 v.3.3.2, ggpubr v.0.4.0 were used in this study. The python packages used in this study include Pandas v. 1.2.4, Numpy v. 1.20.2, Sklearn v. 0.24.2, Pytorch v. 1.9.0+cu102, torch_lr_finder v. 0.2.1, statsmodels v. 0.12.2, Matplotlib v. 3.4.1, pROC v 1.17.0.1, rpy2 v 3.4.5. The models were run on Yale's Milgram high performance computing cluster. Descriptive statistics used the Kruskal-Wallis and Wilcoxon Rank sum tests for continuous values, chi-square for categorical values, two-proportions Z-test for proportions between groups. Statistical analyses were performed using R [1], version 3.6.3 and Python, version 3.9.13. Univariate and multivariate survival analyses were performed with the R survival package v.3.5-3.

References

1. Ripley BD. The R project in statistical computing. *MSOR Connections. The newsletter of the LTSN Maths* [Internet] Citeseer; 2001; Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.430.3979&rep=rep1&type=pdf.

Classifier Code

```
import numpy as np
import pandas as pd
from sklearn.datasets import load_iris
import xgboost as xgb
from sklearn.model_selection import GridSearchCV, cross_val_score, train_test_split
from sklearn.metrics import accuracy_score, classification_report
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.metrics import accuracy_score, precision_score, recall_score, confusion_matrix,
f1_score, roc_auc_score
import shap

#import joblib
+*In[ ]:*+
[source, ipython3]
----
import seaborn as sns
----


+*In[ ]:*+
[source, ipython3]
----
from sklearn.metrics import roc_curve, auc
import matplotlib.pyplot as plt
from sklearn.model_selection import GridSearchCV, cross_val_score, train_test_split
from sklearn.calibration import calibration_curve
from sklearn.metrics import precision_recall_curve
from sklearn.model_selection import cross_val_score
from hyperopt import STATUS_OK
from hyperopt import hp
----
```

For this code, we are trying to build up a classifier based on the
previous unsupervised learning clustering outcome. We try to apply this
classifier to similar set of patient’s data but with different criteria
for admission (covid).

1. Cross-validation for 5-fold to find the best number of rounds for an
80-20 split each time, and then select the best numbers and train the
entire dataset.

```
+*In[2]:*+
[source, ipython3]
----
# loading the cluster_data
df = pd.read_csv('cluster_res.csv')#Original dataset used to identify the subgroups
df = df.drop(columns=['Unnamed: 0'])
----
```

```
+*In[3]:*+
[source, ipython3]
----
df.columns
----
```

```
+*Out[3]:*+
----Index(['Age', 'Sex', 'BMI', 'BP_Systolic', 'BP_Diastolic', 'MAP',
       'Smoking.Status.Simple', 'Insurance.Simple', 'CAD', 'Cereb_VD', 'DM',
       'CKD', 'CHF', 'Allergic.Rhinitis', 'GERD', 'Nasal.Polyposis',
       'Lung.Cancer', 'Sleep.Apnea', 'HTN', 'ALBUTEROL', 'ICS_LABA', 'LAMA',
       'ANTIBIOTIC', 'IPRATROPIUM', 'SYSTEMIC_STEROID', 'FLU_VACCINE', 'ICS',
       'LTRA', 'THEOPHYLLINE', 'PNEUMOCOCCAL_VACCINE', 'NICOTINE',
       'OSELTAMIVIR', 'WBC', 'Neutrophils', 'Eosinophils', 'Basophils',
       'Monocytes', 'Lymphocytes', 'Abs.Neu.Count', 'Abs.Eo.Count',
       'Abs.Baso.Count', 'Abs.Mono.Count', 'Abs.Lymph.Count', 'RBCC',
       'Hematocrit', 'Hemoglobin', 'MCH', 'MCHC', 'MCV', 'MPV', 'Platelets',
       'cluster', 'Study.ID'],
      dtype='object')----
```

= Method 1

= Data-processing

```
+*In[4]:*+
[source, ipython3]
----
# check the data condition to see if there is any missing data and types
# XGB won't able to processing objective
# so need to convert to dummy
```

```
df.dtypes
----
```

```
+*Out[4]:*+
----Age              int64
Sex               object
BMI              float64
BP_Systolic          int64
BP_Diastolic         int64
MAP               float64
Smoking.Status.Simple    object
Insurance.Simple       object
CAD               object
Cereb_VD             object
DM                object
CKD               object
CHF               object
Allergic.Rhinitis   object
GERD                object
Nasal.Polyposis       object
Lung.Cancer          object
Sleep.Apnea          object
HTN               object
ALBUTEROL              object
ICS_LABA             object
LAMA               object
ANTIBIOTIC            object
IPRATROPIUM             object
SYSTEMIC_STEROID            object
FLU_VACCINE             object
ICS               object
LTRA               object
THEOPHYLLINE             object
PNEUMOCOCCAL_VACCINE             object
NICOTINE             object
OSELTAMIVIR             object
WBC               float64
Neutrophils        float64
Eosinophils         float64
Basophils          float64
Monocytes           float64
Lymphocytes          float64
Abs.Neu.Count         float64
Abs.Eo.Count         float64
Abs.Baso.Count        float64
Abs.Mono.Count         float64
Abs.Lymph.Count        float64
RBCC               float64
Hematocrit          float64
Hemoglobin           float64
```

```
MCH                      float64
MCHC                     float64
MCV                      float64
MPV                      float64
Platelets                  int64
cluster                    int64
Study.ID                  object
dtype: object----
```

```
+*In[5]:*+
[source, ipython3]
----
df.isnull().sum()
----
```

```
+*Out[5]:*+
----Age                      0
Sex                      0
BMI                      0
BP_Systolic              0
BP_Diastolic             0
MAP                      0
Smoking.Status.Simple    0
Insurance.Simple         0
CAD                      0
Cereb_VD                 0
DM                       0
CKD                      0
CHF                      0
Allergic.Rhinitis        0
GERD                     0
Nasal.Polyposis          0
Lung.Cancer              0
Sleep.Apnea              0
HTN                      0
ALBUTEROL                0
ICS_LABA                 0
LAMA                     0
ANTIBIOTIC               0
IPRATROPIUM              0
SYSTEMIC_STEROID         0
FLU_VACCINE              0
ICS                      0
LTRA                     0
THEOPHYLLINE             0
PNEUMOCOCCAL_VACCINE     0
NICOTINE                 0
OSELTAMIVIR              0
WBC                      0
```

Published online October 17, 2024 https://doi.org/10.15326/jcopdf.2024.0556

```
Neutrophils          0
Eosinophils          0
Basophils            0
Monocytes            0
Lymphocytes          0
Abs.Neu.Count        0
Abs.Eo.Count         0
Abs.Baso.Count       0
Abs.Mono.Count       0
Abs.Lymph.Count      0
RBCC                 0
Hematocrit           0
Hemoglobin           0
MCH                  0
MCHC                 0
MCV                  0
MPV                  0
Platelets            0
cluster              0
Study.ID             0
dtype: int64----


+*In[ ]:*+
[source, ipython3]
----
#covid_data['Smoking.Status.Simple'] = np.nan
#covid_data['Insurance.Simple'] = np.nan
#covid_data['Allergic.Rhinitis'] = np.nan
#covid_data['Nasal.Polyposis'] = np.nan
#covid_data['Sleep.Apnea'] = np.nan
#covid_data['ICS_LABA'] = np.nan
#covid_data['FLU_VACCINE'] = np.nan
#covid_data['PNEUMOCOCCAL_VACCINE'] = np.nan
----


+*In[5]:*+
[source, ipython3]
----
col_to_drop = ['Lung.Cancer', 'FLU_VACCINE', 'Allergic.Rhinitis','ALBUTEROL', 'LTRA',
'IPRATROPIUM','Nasal.Polyposis',
            'THEOPHYLLINE', 'OSELTAMIVIR',
'cluster','Study.ID','Smoking.Status.Simple','Insurance.Simple',
            'Allergic.Rhinitis', 'Nasal.Polyposis', 'Sleep.Apnea','ICS_LABA','FLU_VACCINE',
'PNEUMOCOCCAL_VACCINE']
----


+*In[6]:*+
[source, ipython3]
```

```
----
# Identify the target column and separate it from the features
X = df.drop(col_to_drop, axis=1)
y = df['cluster']
----


+*In[7]:*+
[source, ipython3]
----
# Convert the string varaibles to one hot encoding
categorical_columns = list(X.select_dtypes(include=['object']).columns)
label_encoders = {}
for col in categorical_columns:
    label_encoders[col] = LabelEncoder()
    X[col] = label_encoders[col].fit_transform(X[col])
----


+*In[26]:*+
[source, ipython3]
----
X.columns
----


+*Out[26]:*+
----Index(['Age', 'Sex', 'BMI', 'BP_Systolic', 'BP_Diastolic', 'MAP', 'CAD',
       'Cereb_VD', 'DM', 'CKD', 'CHF', 'GERD', 'HTN', 'LAMA', 'ANTIBIOTIC',
       'SYSTEMIC_STEROID', 'ICS', 'NICOTINE', 'WBC', 'Neutrophils',
       'Eosinophils', 'Basophils', 'Monocytes', 'Lymphocytes', 'Abs.Neu.Count',
       'Abs.Eo.Count', 'Abs.Baso.Count', 'Abs.Mono.Count', 'Abs.Lymph.Count',
       'RBCC', 'Hematocrit', 'Hemoglobin', 'MCH', 'MCHC', 'MCV', 'MPV',
       'Platelets'],
      dtype='object')----


+*In[10]:*+
[source, ipython3]
----
X.isnull().sum()
----


+*Out[10]:*+
----Age                 0
Sex                 0
BMI                 0
BP_Systolic         0
BP_Diastolic        0
MAP                 0
```

```
CAD               0
Cereb_VD          0
DM                0
CKD               0
CHF               0
GERD              0
HTN               0
LAMA              0
ANTIBIOTIC        0
SYSTEMIC_STEROID    0
ICS               0
NICOTINE          0
WBC               0
Neutrophils       0
Eosinophils       0
Basophils         0
Monocytes         0
Lymphocytes       0
Abs.Neu.Count     0
Abs.Eo.Count      0
Abs.Baso.Count    0
Abs.Mono.Count    0
Abs.Lymph.Count   0
RBCC              0
Hematocrit        0
Hemoglobin        0
MCH               0
MCHC              0
MCV               0
MPV               0
Platelets         0
dtype: int64----
```

```
+*In[11]:*+
[source, ipython3]
----
# Split training and test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
----
```

```
+*In[12]:*+
[source, ipython3]
----
# Define a grid of hyperparameters to search
param_grid = {
    'learning_rate': [0.01, 0.1, 0.2],
    'n_estimators': [100, 200, 300],
    'max_depth': [3, 4, 5],
    'min_child_weight': [1, 2, 3],
    'gamma': [0, 0.1, 0.2],
    'subsample': [0.8, 0.9, 1.0],
    'colsample_bytree': [0.8, 0.9, 1.0],
}
----
```

```
+*In[13]:*+
[source, ipython3]
----
xgb_model = xgb.XGBClassifier(objective='multi:softprob')
----
```

```
+*In[14]:*+
[source, ipython3]
----
# Perform Grid Search with cross-validation
grid_search = GridSearchCV(estimator=xgb_model, param_grid=param_grid,
scoring='precision', cv=5, n_jobs=-1)
grid_search.fit(X_train, y_train)
----
```

```
+*Out[14]:*+
GridSearchCV(cv=5,
             estimator=XGBClassifier(base_score=None, booster=None,
                                     callbacks=None, colsample_bylevel=None,
                                     colsample_bynode=None,
                                     colsample_bytree=None, device=None,
                                     early_stopping_rounds=None,
                                     enable_categorical=False, eval_metric=None,
                                     feature_types=None, gamma=None,
                                     grow_policy=None, importance_type=None,
                                     interaction_constraints=None,
                                     learning_rate=None,...
                                     missing=nan, monotone_constraints=None,
                                     multi_strategy=None, n_estimators=None,
                                     n_jobs=None, num_parallel_tree=None,
                                     objective='multi:softprob', ...),
             n_jobs=-1,
             param_grid={'colsample_bytree': [0.8, 0.9, 1.0],
                         'gamma': [0, 0.1, 0.2],
                         'learning_rate': [0.01, 0.1, 0.2],
                         'max_depth': [3, 4, 5], 'min_child_weight': [1, 2, 3],
                         'n_estimators': [100, 200, 300],
                         'subsample': [0.8, 0.9, 1.0]},
             scoring='precision')----
```

```
+*In[15]:*+
[source, ipython3]
----
# Get the best hyperparameters
best_params = grid_search.best_params_
print(f"Best Hyperparameters: {best_params}")
----


+*Out[15]:*+
----
Best Hyperparameters: {'colsample_bytree': 0.8, 'gamma': 0, 'learning_rate': 0.01, 'max_depth':
3, 'min_child_weight': 1, 'n_estimators': 100, 'subsample': 0.8}
----


+*In[27]:*+
[source, ipython3]
----
# Train the final model with the best hyperparameters using the full training dataset
best_xgb_model = xgb.XGBClassifier(**best_params, objective='multi:softprob')
best_xgb_model.fit(X_train, y_train)
----


+*Out[27]:*+
----XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=0.8, device=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric=None, feature_types=None,
              gamma=0, grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=0.01, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=3, max_leaves=None,
              min_child_weight=1, missing=nan, monotone_constraints=None,
              multi_strategy=None, n_estimators=100, n_jobs=None,
              num_parallel_tree=None, objective='multi:softprob', ...)----


+*In[28]:*+
[source, ipython3]
----
#cross-validation
cv_scores = cross_val_score(best_xgb_model, X, y, cv=5)
----


+*In[29]:*+
[source, ipython3]
----
print(f'Cross-validation scores: {cv_scores}')
print(f'Mean CV score: {np.mean(cv_scores)}')
```

```
----


+*Out[29]:*+
----
Cross-validation scores: [0.83333333 0.79827089 0.78097983 0.81268012 0.76368876]
Mean CV score: 0.7977905859750241
----


+*In[30]:*+
[source, ipython3]
----
# Predict using the final model on the test set
y_pred = best_xgb_model.predict(X_test)
----


+*In[31]:*+
[source, ipython3]
----
# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy on test set: {accuracy}')

# Calculate confusion matrix
confusion = confusion_matrix(y_test, y_pred)
print(f'Confusion Matrix:\n{confusion}')
----


+*Out[31]:*+
----
Accuracy on test set: 0.7902298850574713
Confusion Matrix:
[[159  16   3]
 [ 27  90   5]
 [ 11  11  26]]
----


+*In[40]:*+
[source, ipython3]
----
# Create a heatmap using seaborn
sns.heatmap(confusion, annot=True, fmt='d', xticklabels=['Predicted class 1', 'Predicted class 2',
'Predicted class 3'], yticklabels=['Actual class 1', 'Actual class 2', 'Actual class 3'], cmap='Blues')
----


+*Out[40]:*+
----<AxesSubplot:>
![png](output_26_1.png)
----


+*In[32]:*+
[source, ipython3]
----
# Calculate precision
precision = precision_score(y_test, y_pred, average='weighted')
print(f'Precision: {precision}')

# Calculate recall
recall = recall_score(y_test, y_pred, average='weighted')
print(f'Recall: {recall}')
----


+*Out[32]:*+
----
Precision: 0.7879799161447826
Recall: 0.7902298850574713
----


+*In[33]:*+
[source, ipython3]
----
# Calculate f1
f1 = f1_score(y_test, y_pred, average='weighted')
print(f'f1: {f1}')
----


+*Out[33]:*+
----
f1: 0.785246751662438
----


+*In[34]:*+
[source, ipython3]
----
# Visualize feature importances using SHAP
# Create an Explanation object for SHAP values
explainer = shap.Explainer(best_xgb_model, X_train)
shap_values = explainer.shap_values(X_test)
----


+*Out[34]:*+
----
[15:45:33] WARNING: /Users/runner/work/xgboost/xgboost/src/c_api/c_api.cc:1240: Saving into
deprecated binary model format, please consider using `json` or `ubj`. Model format will default
to JSON in XGBoost 2.2 if not specified.
----
```

```
+*In[35]:*+
[source, ipython3]
----
shap.summary_plot(shap_values, X_test, feature_names=X.columns, plot_type='bar',
max_display=X.shape[1])
----
```

```
+*Out[35]:*+
----
![png](output_30_0.png)
----
```

```
+*In[ ]:*+
[source, ipython3]
----
joblib.dump(best_xgb_model, 'best_xgboost_model_v2.pkl')
----
```

Supplementary Table

Cytokine	Reference	Cardio-Renal	Eosinophilic	p-value	FDR
IL-1b	1 (1-5)	1 (1-5)	5 (1-5)	0.006	0.048
IL-2	1 (1-5)	1 (1-5)	2.1 (1-5)	0.057	0.125
IL-2R	2288 (1357-3537)	2390 (1394-4300)	1562 (1057-2440)	0.013	0.048
IL-4	5 (5)	5 (5)	5 (5)	0.572	0.699
IL-5	5 (5)	5 (5)	5 (5)	0.773	0.850
IL-6	18.7 (6.4-56.3)	24.1 (7.2-68.3)	16 (5-36.5)	0.050	0.125
IL-8	22.7 (5-44.2)	23.4 (5-45.7)	5 (5-25.2)	0.012	0.048
IL-10	10.9 (6.12-20.8)	11.4 (6.35-22)	9.5 (5-19.1)	0.260	0.358
IL-12	5 (5)	5 (5)	5 (5)	0.254	0.358
IL-13	5 (5)	5 (5)	5 (5)	0.239	0.358
IL-17	5 (5)	5 (5)	5 (5)	0.867	0.867

Values are median (IQR)

Published online October 17, 2024 https://doi.org/10.15326/jcopdf.2024.0556

STROBE Statement—Checklist of items that should be included in reports of ***cohort studies***

	Item No	Recommendation	Page No
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract (*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	1 2-3
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	4-5
Objectives	3	State specific objectives, including any prespecified hypotheses	4-5
Meth			
Study design	4	Present key elements of study design early in the paper	6-7 Supp
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	6-7 Supp
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up (*b*) For matched studies, give matching criteria and number of exposed and unexposed	6-7 Supp
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	6-7 Supp
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	6-7 Supp
Bias	9	Describe any efforts to address potential sources of bias	6-7 Supp
Study size	10	Explain how the study size was arrived at	NA
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	6-7 Supp
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	6-7 Supp

		(*b*) Describe any methods used to examine subgroups and interactions (*c*) Explain how missing data were addressed (*d*) If applicable, explain how loss to follow-up was addressed (*e*) Describe any sensitivity analyses	Table 3
Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed (b) Give reasons for non-participation at each stage (c) Consider use of a flow diagram	8-12 Tables 1 and 4
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders (b) Indicate number of participants with missing data for each variable of interest (c) Summarise follow-up time (eg, average and total amount)	8-12 Tables 1 and 4
Outcome data	15*	Report numbers of outcome events or summary measures over time	8-12 Tables 2 and 5

Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included (*b*) Report category boundaries when continuous variables were categorized (*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	8-12 Tables
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	Table 3, figure 2
Discussion			
Key results	18	Summarise key results with reference to study objectives	14-17
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	14-17
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	14-17
Generalisability	21	Discuss the generalisability (external validity) of the study results	14-17
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	Title page

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.