Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

CHEST

Phenotypes and Subphenotypes of Patients With COVID-19

A Latent Class Modeling Analysis

Xiaofeng Wang, PhD; Lara Jehi, MD; Xinge Ji, MS; and Peter J. Mazzone, MD, MPH, FCCP

BACKGROUND: Since COVID-19 was identified, its clinical and biological heterogeneity has been recognized. Identifying COVID-19 phenotypes might help guide basic, clinical, and translational research efforts.

RESEARCH QUESTION: Does the clinical spectrum of patients with COVID-19 contain distinct phenotypes and subphenotypes?

STUDY DESIGN AND METHODS: We included adult patients (≥ 18 years) positive for laboratory-confirmed SARS-CoV-2 infection from a prospective COVID-19 registry database in the Cleveland Clinic Health System in Ohio and Florida. The patients were split into training and testing sets. Using latent class analysis (LCA), we first identified phenotypic clusters of patients with COVID-19 based on demographics, comorbidities, and presenting symptoms. We then identified subphenotypes of hospitalized patients with additional blood biomarker data measured on hospital admission. The associations of phenotypes/subphenotypes and clinical outcomes were investigated. Multivariable prediction models were established to predict assignment to the LCA-defined phenotypes and subphenotypes and then evaluated on an independent testing set.

RESULTS: We analyzed data for 20,572 patients. Seven phenotypes were identified on the basis of different profiles of presenting COVID-19 symptoms and existing comorbidities, including the following groups: young, no symptoms; young, symptoms; middle-aged, no symptoms; middle-aged, symptoms; middle-aged, comorbidities; old, no symptoms; and old, symptoms. The rates of inpatient hospitalization for the phenotypes were significantly different ($P <$ .001). Five subphenotypes were identified for the subgroup of hospitalized patients, including the following subgroups: young, elevated WBC and platelet counts; middle-aged, lymphopenic with elevated C-reactive protein; middle-aged, hyperinflammatory; old, leukopenic with comorbidities; and old, hyperinflammatory with kidney dysfunction. The hospital mortality and the times from hospitalization to ICU transfer or death were significantly different ($P <$ .001). The models for predicting the LCA-defined phenotypes and subphenotypes showed high discrimination (concordance index, 0.92 and 0.91).

INTERPRETATION: Hypothesis-free LCA-defined phenotypes and subphenotypes of patients with COVID-19 can be identified. These may help clinical investigators conduct stratified analyses in clinical trials and assist basic science researchers in characterizing the pathobiology of the spectrum of COVID-19 presentations. CHEST 2021; 159(6):2191-2204

KEY WORDS: clinical trials; COVID-19; latent class analysis; phenotypes; subphenotypes

ABBREVIATIONS: BIC = Bayesian information criteria; *C* index = concordance index; CRP = C-reactive protein; LCA = latent class analysis

AFFILIATIONS: From the Department of Quantitative Health Sciences (X. Wang and X. Ji), Lerner Research Institute, Cleveland Clinic, Cleveland, OH; Neurological Institute (L. Jehi), Cleveland Clinic, Cleveland, OH; and Respiratory Institute (P. J. Mazzone), Cleveland Clinic, Cleveland, OH.

FUNDING/SUPPORT: The authors have reported to *CHEST* that no funding was received for this study.

CORRESPONDENCE TO: Xiaofeng Wang, PhD; e-mail: wangx6@ccf.org

DOI: https://doi.org/10.1016/j.chest.2021.01.057

The clinical spectrum of COVID-19 is broad, ranging from asymptomatic infection to severe pneumonia with respiratory failure. Many studies have reported on the clinical characteristics of COVID-19.[1-5] The largest cohort study reported to date included 72,314 cases from the Chinese Center for Disease Control and Prevention.[6] In this study, 81% of patients had mild to moderate disease (mild symptoms up to mild pneumonia), 14% had severe disease (dyspnea, hypoxia, or $>$ 50% lung involvement on imaging), and 5% had critical disease (respiratory failure, shock, or multiorgan system dysfunction). These findings drew early broad brushstrokes of the clinical disease spectrum. A more complex picture later emerged, with symptoms extending beyond the respiratory system, and unpredictable clinical deterioration in patients initially thought to have mild disease. This uncovered a need for a more sophisticated classification mirroring the heterogeneous clinical progression. In parallel, several efforts have evaluated laboratory-based disease biomarkers. Guan et al[5] found that elevated serum alanine aminotransferase and aspartate aminotransferase levels, lactate dehydrogenase, C-reactive protein (CRP), and ferritin levels may be associated with greater illness severity manifest by ARDS and acute kidney injury, resulting in a higher mortality rate.

The clinical and biological heterogeneity of COVID-19[7,8] has led to an incomplete categorization of disease phenotypes. In contrast, research in cancer, ARDS, and asthma has been able to identify disease phenotypes with important therapeutic implications.[9-11] At present, a variety of therapeutic options for COVID-19 are under investigation. Recognizing different phenotypes and subphenotypes of COVID-19 might help guide basic, clinical, and translational research efforts. Such an understanding could help clinicians and researchers stratify patients for clinical trials and customize therapy.

Latent class analysis (LCA), a subset of structural equation modeling, is a well-validated statistical technique for identifying unmeasured class membership among subjects, using categorical and/or continuous observed variables.[12] It uses mixture modeling to find the best-fitting model under the assumption that individuals can be divided into subgroups based on an unobservable construct. The subgroups are called latent classes. LCA is an unsupervised analysis in that it asks whether there are subgroups of individuals defined by a combination of variables, without mandating consideration of an outcome. LCA has been successfully used in respiratory and critical care medicine, for example, in the identification of phenotypes of childhood asthma[9] and subphenotypes of ARDS.[11] Despite widespread recognition of the heterogeneity within COVID-19, little work has robustly studied if/what subphenotypes exist. This article aims to take advantage of the wealth of clinical and blood biomarker data available from the prospective COVID-19 Registry of the Cleveland Clinic Health System[13] by using latent class modeling approaches to identify phenotypes and subphenotypes of patients with COVID-19 and to test their association with clinical outcomes.

Methods

Study Design and Population

A prospective COVID-19 registry database was set up in March 2020 to align data collection for research with clinical care of all patients who are tested for COVID-19 in the Cleveland Clinic Health System. Data capture was facilitated by creating standardized clinical templates that are implemented across the health care system.[13] Study data were collected and managed using Research Electronic Data Capture (REDCap; Vanderbilt University) tools hosted at the Cleveland Clinic. Registry variables were chosen to reflect the available literature on COVID-19 disease characterization, progression, and treatment. The study was approved by the Cleveland Clinic Institutional Review Board. The requirement for written informed consent was waived.

We included adult patients (age, $\geq$ 18 years) with confirmed SARS-CoV-2 infection in the Cleveland Clinic Health System in the United States between March 12 and October 31, 2020. COVID-19 was confirmed by reverse transcription-polymerase chain reaction for SARS-CoV-2. The testing protocols were previously described.[13]

Data Extraction

Patient demographics, comorbidities, presenting symptoms, and medications were retrieved and analyzed. Data on comorbidities including cancer, COPD/emphysema, asthma, diabetes mellitus, hypertension, coronary artery disease, transplantation, multiple sclerosis, inflammatory bowel disease, and immunosuppressive disease were extracted from the electronic health record (Epic; Epic Systems Corporation). Immunosuppressive disease was defined on the basis of the Agency for Healthcare Research and Quality patient safety indicator (Appendix I: Immunocompromised State Diagnosis and Procedure Codes[14]).

For hospitalized patients, routine blood examination results from hospital admission were extracted, including CBC count, basic metabolic panel, coagulation profile, and renal and liver function tests. We excluded blood biomarkers if they were missing for

≥ 30% of subjects. We finally selected 17 candidate blood biomarkers based on the current literature.[1-6] They represented a diverse range of biological processes, including absolute lymphocyte count, absolute neutrophil count, albumin, alanine aminotransferase, alkaline phosphatase, BUN, chloride, CRP, creatinine, ferritin, hematocrit, hemoglobin, platelets, potassium, RBC distribution width, total bilirubin, and WBC count.

Clinical Outcome Measures

For all patients who tested positive for SARS-CoV-2, the primary outcome measured was the need for inpatient hospitalization. For the subgroup cohort of hospitalized patients, the primary outcome measured was a composite of transfer to an ICU or in-hospital mortality (World Health Organization [WHO] COVID-19 Ordinal Scale for Clinical Improvement, scores 6-8) vs no transfer to ICU and alive (WHO Ordinal Scale for Clinical Improvement, scores 3-5) during hospitalization. The secondary outcome for hospitalized patients was in-hospital mortality only. Patients still hospitalized at the time of analysis were censored. The outcomes for all patients were followed up until December 5, 2020.

Statistical Analysis

The study patients were split into training and testing sets based on the time their information was entered into the database. Patients whose data were collected before August 28, 2020, were considered the training sample, whereas patients whose data were collected between August 29 and October 31, 2020, were considered the testing sample. The study variables were described using sample median with interquartile range or number with proportion. Categorical variables were compared using the Pearson χ^2 test or Fisher exact test, whereas continuous variables were compared using the Mann-Whitney *U* test. We developed two latent class models in this study: the first one was for all patients who tested positive; the second one was for hospitalized patients only.

The candidate class-defining variables in the first model included demographics, comorbidities, and presenting symptoms. BMI was excluded as it was missing for 47% of the cohort. The candidate class-defining variables in the second model included five demographic/clinical variables (age, sex, the presence of cancer, the presence of COPD/emphysema, and the number of other comorbidities), and the 17 blood biomarker measures at hospital admission (described in the section "Data Extraction"). The comorbidities selected were based on available literature showing that patients infected with SARS-CoV-2 and who had cancer or COPD had poor outcomes with a high occurrence of clinically severe events and mortality.[15,16]

Missing value imputation for biomarkers was compared using median imputation and multivariate imputation by chained equations, and the median imputation was used in the analysis. Log-transformation was performed on continuous variables, if needed, to reduce or remove the skewness of the original data. Standardization was then processed to make all continuous variables fit on the same scale.

Latent class variable selection analysis was performed on the basis of the Fop et al[17] method, because removing unnecessary variables and parameters can improve classification performance and the precision of parameter estimates. Parameters of the latent class models were estimated using expectation-maximization methods.[18] Latent classes were determined without consideration of clinical outcomes. The best-fitting models were determined on the basis of the Bayesian information criterion (BIC). Individuals were then assigned to the class for which they had the highest posterior probability of belonging.

Kruskal-Wallis, Pearson χ^2, and Fisher exact tests, as appropriate, were performed to compare the clinical, laboratory, and outcome variables between the resulting classes. Kaplan-Meier survival analysis was conducted to describe survival characteristics among the different subclasses in hospitalized patients. Finally, we built multinomial logistic regression models by the bias correction method[19,20] to predict latent class membership. Bootstrap internal validation was conducted to assess the discriminative ability of the models.

All analyses were performed with the R software program (version 3.6.3; R Foundation for Statistical Computing) and SAS software (version 9.4; SAS Institute). The level of statistical significance was set at $P < .05$ (two-tailed).

Results

Patient Characteristics

There were 285,783 patients who presented with symptoms of respiratory tract infection or had close contact with someone with confirmed COVID-19 from March 12 to October 31, 2020, and who underwent SARS-CoV-2 testing. Of the 21,978 patients who tested positive for SARS-CoV-2, 20,572 adult patients served as the study population (Fig 1). The patients were split into training and testing sets, with 11,818 patients included in the model training set and 8,754 patients serving as the testing cohort. A total of 3,546 patients (2,655 and 893 patients in the training and testing sets, respectively) were admitted to a hospital in the Cleveland Clinic Health System, which served as the subpopulation in the subgroup analysis. Baseline demographic and clinical characteristics of the training and testing cohorts in the study are presented in e-Tables 1 through 4.

Characteristics of LCA-Defined Phenotypes for Patients Who Tested Positive

The best-fitting model, as indicated by the BIC, was a seven-class model using one demographic variable, six symptom variables, and seven comorbidity variables: the demographic variable was age; the symptom variables included the presences of cough, fever, fatigue, sputum production, shortness of breath, and diarrhea; the comorbidity variables included the presence of asthma, COPD/emphysema, diabetes, hypertension, coronary artery disease, heart failure, cancer, and immunosuppressive disease. Table 1 displays the group comparison results to help us understand the clinical characteristics that distinguished each phenotype.

Phenotype class 1 (young, no symptoms; 32% of the sample) and phenotype class 2 (young, symptoms; 14% of the sample) contained mostly young adults, with median ages 33 and 34 years, respectively. Patients in

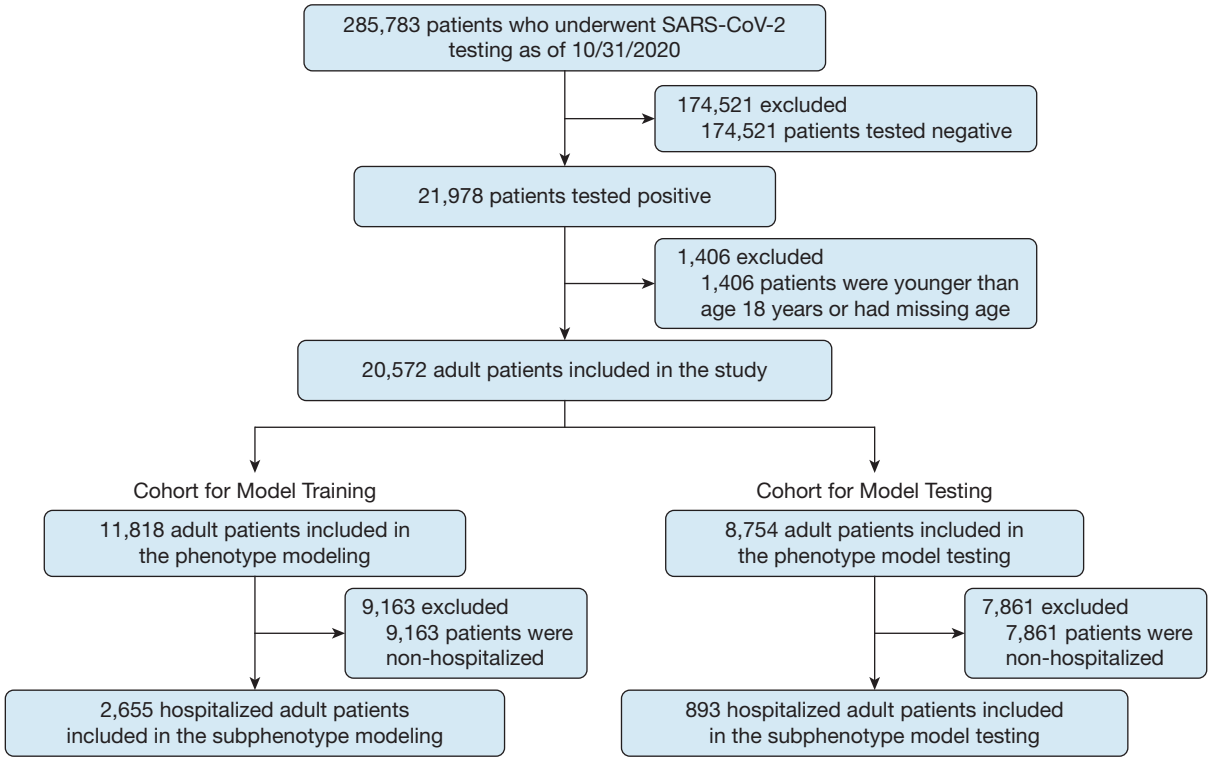

Figure 1 – *Study flow diagram.*

class 1 had very few COVID-19 symptoms and comorbidities. Patients in class 2 also had few comorbidities but did have typical COVID-19 symptoms, such as cough (85%) and fever (65%).

Phenotype class 3 (middle-aged, symptoms; 8% of the sample), phenotype class 4 (middle-aged, no symptoms; 21% of the sample), and phenotype class 5 (middle-aged, comorbidities; 11% of the sample) were more heavily populated with individuals with median ages 56, 61, and 59 years, respectively. Patients in class 3 had almost all of the common COVID-19 symptoms, whereas patients in class 4 had nearly no COVID-19 symptoms. In contrast, in class 5 some symptoms were common (cough, 82%; fever, 61%; fatigue, 50%), but some were not (sputum production, 27%; shortness of breath, 38%; diarrhea, 20%, vomiting, 13%). Comorbidities were more common in the three middle-aged groups than in the young adult classes. The frequencies of hypertension, diabetes, and cancer in class 4 and class 5 were much higher than in class 3.

Phenotype class 6 (old, no symptoms; 8% of the sample) and phenotype class 7 (old, symptoms; 6% of the sample) were populated with individuals with median ages 74 and 76 years, respectively. Patients in class 6 had few COVID-19 symptoms, whereas those in class 7 had almost all of the common symptoms. Both classes had high frequencies of comorbidities.

The χ^2 analysis showed a significant association between the LCA-defined phenotypes and inpatient hospitalization ($P < .001$). The rates of inpatient hospitalization for the seven phenotypes were 9.7%, 9.8%, 22%, 24%, 33%, 47%, and 59%, respectively.

Characteristics of LCA-Defined Subphenotypes for Hospitalized Patients

The best-fitting model, as indicated by the BIC, was a five-class model using four clinical variables and seven blood biomarkers. The clinical variables included age, the presence of cancer, the presence of COPD/emphysema, and the number of other comorbidities, and the blood biomarkers included WBC count, lymphocyte count, CRP, creatinine, albumin, platelet count, and hemoglobin. Table 2 displays the class comparison results to illustrate the clinical and biological characteristics that distinguished each subphenotype. Figure 2 shows the latent profile plots of the subphenotypes identified.

Subphenotype Class 1 (Young, Elevated WBC and Platelet Counts): This subclass included 14% of the sample. It represented the control COVID-19 clinical

TABLE 1] Comparison of Phenotypes of Adult Patients With Confirmed COVID-19 Identified by Latent Class Analysis

Variable	Overall (N = 11,818)	Class 1 (n = 3,803)	Class 2 (n = 1,594)	Class 3 (n = 988)	Class 4 (n = 2,503)	Class 5 (n = 1,310)	Class 6 (n = 926)	Class 7 (n = 694)	*P* Value
Demographics									
Age, median (IQR), y	50 (33-65)	33 (26-44)	34 (26-44)	56 (45-66)	61 (52-71)	59 (50-67)	74 (66-83)	76 (66-86)	< .001
Sex, No. (%)									< .001
Female	6,500 (55)	2,202 (58)	887 (56)	498 (50)	1,402 (56)	713 (54)	454 (49)	344 (50)	
Male	5,318 (45)	1,601 (42)	707 (44)	490 (50)	1,101 (44)	597 (46)	472 (51)	350 (50)	
Race, No. (%)									< .001
White	6,484 (55)	1,919 (50)	808 (51)	569 (58)	1,470 (59)	684 (52)	587 (63)	447 (64)	
Black	3,583 (30)	1,156 (30)	468 (29)	251 (25)	745 (30)	470 (36)	282 (30)	211 (30)	
Other	1,751 (15)	728 (19)	318 (20)	168 (17)	288 (12)	156 (12)	57 (6.2)	36 (5.2)	
Ethnicity, No. (%)									< .001
Hispanic	1,440 (12)	574 (15)	236 (15)	139 (14)	278 (11)	144 (11)	44 (4.8)	25 (3.6)	
Non-Hispanic	9,231 (78)	2,727 (72)	1,137 (71)	721 (73)	2,056 (82)	1,082 (83)	858 (93)	650 (94)	
Unknown	1,147 (9.7)	502 (13)	221 (14)	128 (13)	169 (6.8)	84 (6.4)	24 (2.6)	19 (2.7)	
Smoking, No. (%)									< .001
Current smoker	874 (7.4)	277 (7.3)	121 (7.6)	60 (6.1)	185 (7.4)	103 (7.9)	70 (7.6)	58 (8.4)	
Former smoker	2,834 (24)	414 (11)	283 (18)	309 (31)	624 (25)	405 (31)	466 (50)	333 (48)	
Nonsmoker	6,139 (52)	2,063 (54)	1,004 (63)	444 (45)	1,267 (51)	739 (56)	348 (38)	274 (39)	
Unknown	1,971 (17)	1,049 (28)	186 (12)	175 (18)	427 (17)	63 (4.8)	42 (4.5)	29 (4.2)	
Presenting symptom, No. (%)									
Cough	4,164 (35)	139 (3.7)	1,357 (85)	958 (97)	0 (0)	1,077 (82)	28 (3.0)	605 (87)	< .001
Fever	3,286 (28)	67 (1.8)	1,032 (65)	904 (91)	1 (< 0.1)	801 (61)	33 (3.6)	448 (65)	< .001
Fatigue	3,334 (28)	43 (1.1)	1,015 (64)	988 (100)	14 (0.6)	653 (50)	26 (2.8)	595 (86)	< .001
Sputum production	2,381 (20)	0 (0)	653 (41)	894 (90)	0 (0)	351 (27)	0 (0)	483 (70)	< .001
Flu-like symptoms	3,918 (33)	202 (5.3)	1,276 (80)	951 (96)	22 (0.9)	907 (69)	18 (1.9)	542 (78)	< .001
Shortness of breath	2,682 (23)	11 (0.3)	715 (45)	835 (85)	0 (0)	494 (38)	49 (5.3)	578 (83)	< .001
Diarrhea	2,178 (18)	20 (0.5)	573 (36)	876 (89)	2 (< 0.1)	265 (20)	15 (1.6)	427 (62)	< .001
Vomiting	1,546 (13)	34 (0.9)	340 (21)	689 (70)	0 (0)	168 (13)	7 (0.8)	308 (44)	< .001
Comorbidities, No. (%)									
Asthma	1,727 (15)	471 (12)	225 (14)	119 (12)	305 (12)	219 (17)	241 (26)	147 (21)	< .001
COPD/emphysema	725 (6.1)	12 (0.3)	0 (0)	32 (3.2)	100 (4.0)	45 (3.4)	310 (33)	226 (33)	< .001
Diabetes	2,108 (18)	44 (1.2)	18 (1.1)	138 (14)	570 (23)	452 (35)	527 (57)	359 (52)	< .001

(Continued)

TABLE 1] *(Continued)*

Variable	Overall (N = 11,818)	Class 1 (n = 3,803)	Class 2 (n = 1,594)	Class 3 (n = 988)	Class 4 (n = 2,503)	Class 5 (n = 1,310)	Class 6 (n = 926)	Class 7 (n = 694)	*P* Value
Hypertension	4,638 (39)	73 (1.9)	1 (< 0.1)	384 (39)	1,607 (64)	1,014 (77)	898 (97)	661 (95)	< .001
Coronary artery disease	1,120 (9.5)	0 (0)	0 (0)	13 (1.3)	116 (4.6)	82 (6.3)	543 (59)	366 (53)	< .001
Heart failure	879 (7.4)	10 (0.3)	0 (0)	1 (0.1)	32 (1.3)	36 (2.7)	458 (49)	342 (49)	< .001
Cancer	1,208 (10)	35 (0.9)	12 (0.8)	78 (7.9)	377 (15)	192 (15)	297 (32)	217 (31)	< .001
Transplant history	90 (0.8)	2 (< 0.1)	3 (0.2)	5 (0.5)	17 (0.7)	11 (0.8)	32 (3.5)	20 (2.9)	< .001
Multiple sclerosis	78 (0.7)	15 (0.4)	8 (0.5)	7 (0.7)	20 (0.8)	15 (1.1)	6 (0.6)	7 (1.0)	.075
Connective tissue disease	540 (4.6)	30 (0.8)	45 (2.8)	55 (5.6)	75 (3.0)	139 (11)	87 (9.4)	109 (16)	< .001
Inflammatory bowel disease	243 (2.1)	53 (1.4)	35 (2.2)	35 (3.5)	27 (1.1)	39 (3.0)	21 (2.3)	33 (4.8)	< .001
Immunosuppressive disease	1,208 (10)	83 (2.2)	36 (2.3)	61 (6.2)	221 (8.8)	151 (12)	420 (45)	236 (34)	< .001
Outcome, No. (%)									
Hospitalization	2,655 (22)	375 (9.9)	156 (9.8)	221 (22)	612 (24)	434 (33)	448 (48)	409 (59)	< .001

P values are based on Kruskal-Wallis test, Pearson χ^2 test, or Fisher exact test as appropriate. IQR = interquartile range.

and biological response group and was considered our reference group for comparisons. These patients exhibited high platelet counts and high WBC counts. They were the youngest group (median age, 40 years), and many of them exhibited mild anemia. Their CRP, creatinine, and albumin levels were normal. Subclass 1 patients have the lowest rates of ICU transfer (12%) and in-hospital mortality (0%).

Subphenotype Class 2 (Middle-aged, Lymphopenic, With Elevated CRP): This subclass included 21% of the sample, and represented a common inflammatory syndrome group. The median age of patients in this class was 54 years. About 45% of them did not have any comorbidity. Many members of this class had lymphopenia. Compared with subclass 1, patients in this group had moderately elevated CRP, but lower platelet and WBC counts. They had a higher likelihood of being transferred to the ICU (19%). The death rate for this subclass was low (2.8%).

Subphenotype Class 3 (Middle-aged, Hyperinflammatory): This subclass included 20% of the sample, and represented a hyperinflammatory syndrome group. It was characterized by many infection response markers: markedly elevated CRP and platelet and WBC counts compared with subclass 2. Patients in this subclass (median age, 62 years) were significantly older than patients in subclasses 1 and 2, and had more comorbidities than those in subclasses 1 and 2. They also exhibited hypoalbuminemia and lymphopenia. The likelihood of ICU transfer or in-hospital mortality was very high (50%) in this group.

Subphenotype Class 4 (Old, Leukopenic With Comorbidities): Membership in this subclass (25% of the sample) included mostly older patients (median age, 75 years) with multiple comorbidities. They typically exhibited lymphopenia. They had the highest frequency of comorbidities: prior cancer, 31%; COPD/emphysema, 33%; and other comorbidities, 98.8%. The creatinine level (median, 1.13) was higher than in subclasses 1 to 3. They did not have any obvious signs of hyperinflammation. The likelihood of ICU transfer or in-hospital mortality (32%) was lower than for subclass 3.

Subphenotype Class 5 (Old, Hyperinflammatory With Kidney Dysfunction): This subclass included 20% of the sample, and represented a group with kidney dysfunction and hyperinflammatory response. These patients were characterized by renal failure (median creatinine, 2.07 mg/dL), hypoalbuminemia (median

TABLE 2] Comparison of Subphenotypes for Hospitalized Patients With COVID-19, Identified by Latent Class Analysis

Variable	Overall (N = 2,655)	No. of Patients With Missing Data	Subclass 1 (n = 363)	Subclass 2 (n = 568)	Subclass 3 (n = 524)	Subclass 4 (n = 672)	Subclass 5 (n = 5,281)	*P* Value
Demographics								
Age, median (IQR), y	63 (51-75)	0	40 (28-53)	54 (42-62)	62 (53-71)	75 (66-84)	71 (62-80)	< .001
Sex, No. (%)		0						< .001
Female	1,320 (50)		223 (61)	271 (48)	244 (47)	336 (50)	246 (47)	
Male	1,335 (50)		140 (39)	297 (52)	280 (53)	336 (50)	282 (53)	
Race, No. (%)		0						< .001
White	1,368 (52)		141 (39)	293 (52)	277 (53)	372 (55)	285 (54)	
Black	1,083 (41)		188 (52)	208 (37)	200 (38)	266 (40)	221 (42)	
Other	204 (7.7)		34 (9.4)	67 (12)	47 (9.0)	34 (5.1)	22 (4.2)	
Ethnicity, No. (%)		0						< .001
Hispanic	226 (8.5)		34 (9.4)	76 (13)	58 (11)	35 (5.2)	23 (4.4)	
Non-Hispanic	2,388 (90)		321 (88)	486 (86)	451 (86)	631 (94)	499 (95)	
Unknown	41 (1.5)		8 (2.2)	6 (1.1)	15 (2.9)	6 (0.9)	6 (1.1)	
Smoking, No. (%)		0						< .001
Current smoker	192 (7.2)		47 (13)	36 (6.3)	23 (4.4)	38 (5.7)	48 (9.1)	
Former smoker	878 (33)		73 (20)	127 (22)	152 (29)	328 (49)	198 (38)	
Nonsmoker	1,269 (48)		191 (53)	313 (55)	276 (53)	275 (41)	214 (41)	
Unknown	316 (12)		52 (14)	92 (16)	73 (14)	31 (4.6)	68 (13)	
Comorbidities								
COPD/emphysema, No. (%)	347 (13)	0	12 (3.3)	0 (0)	20 (3.8)	207 (31)	108 (20)	< .001
Cancer, No. (%)	429 (16)	0	9 (2.5)	14 (2.5)	46 (8.8)	220 (33)	140 (27)	< .001
No. of other comorbidities, No. (%)		0						< .001
0	648 (24)		182 (50)	257 (45)	137 (26)	8 (1.2)	64 (12)	
1	579 (22)		93 (26)	150 (26)	149 (28)	113 (17)	74 (14)	
2	587 (22)		69 (19)	105 (18)	152 (29)	177 (26)	84 (16)	
3	408 (15)		13 (3.6)	44 (7.7)	62 (12)	173 (26)	116 (22)	
4	228 (8.6)		6 (1.7)	6 (1.1)	20 (3.8)	120 (18)	76 (14)	
5	205 (7.7)		0 (0)	6 (1.1)	4 (0.8)	81 (12)	114 (22)	

(Continued)

TABLE 2] *(Continued)*

Variable	Overall (N = 2,655)	No. of Patients With Missing Data	Subclass 1 (n = 363)	Subclass 2 (n = 568)	Subclass 3 (n = 524)	Subclass 4 (n = 672)	Subclass 5 (n = 5,281)	*P* Value
Home medications								
Nonsteroidal antiinflammatory drugs, No. (%)	837 (32)	0	110 (30)	147 (26)	157 (30)	248 (37)	175 (33)	< .001
Steroids, No. (%)	417 (16)	0	42 (12)	60 (11)	69 (13)	127 (19)	119 (23)	< .001
ACE inhibitor, No. (%)	367 (14)	0	23 (6.3)	64 (11)	80 (15)	129 (19)	71 (13)	< .001
ARB, No. (%)	280 (11)	0	10 (2.8)	39 (6.9)	58 (11)	108 (16)	65 (12)	< .001
Melatonin, No. (%)	155 (5.8)	0	10 (2.8)	15 (2.6)	23 (4.4)	57 (8.5)	50 (9.5)	< .001
Laboratory findings on admission								
Absolute lymphocyte count, median (IQR), K/μL	1.02 (0.71-1.46)	229	1.56 (1.13-2.11)	1.01 (0.81-1.27)	1.00 (0.71-1.38)	0.96 (0.67-1.35)	0.88 (0.52-1.36)	< .001
Absolute neutrophil count, median (IQR), K/μL	4.57 (3.13-6.58)	398	4.12 (2.62-6.44)	3.66 (2.93-4.72)	6.70 (5.50-8.55)	3.67 (2.79-4.75)	5.67 (3.57-8.94)	< .001
Albumin, median (IQR), g/dL	3.70 (3.40-4.00)	207	4.20 (3.90-4.50)	3.90 (3.70-4.10)	3.60 (3.30-3.80)	3.80 (3.50-4.00)	3.30 (2.90-3.70)	< .001
ALT, median (IQR), U/L	24 (15-40)	223	23 (15-40)	28 (18-46)	30 (18-50)	20 (14-31)	20 (12-34)	< .001
Alkaline phosphatase, median (IQR), U/L	74 (59-95)	221	77 (61-98)	67 (53-82)	73 (59-96)	75 (60-94)	82 (62-111)	< .001
BUN, median (IQR), mg/dL	16 (11-27)	140	11 (8-14)	12 (9-16)	16 (11-22)	20 (14-27)	38 (22-56)	< .001
Chloride, median (IQR), mM	99 (96-102)	140	101 (98-103)	98 (96-101)	98 (95-101)	99 (96-103)	99 (95-104)	< .001
CRP, median (IQR), mg/dL	6 (2-12)	551	1 (0-2)	5 (3-8)	13 (9-19)	4 (2-8)	9 (4-16)	< .001
Creatinine, median (IQR), mg/dL	1.03 (0.80-1.41)	140	0.82 (0.67-1.00)	0.90 (0.76-1.10)	0.96 (0.77-1.19)	1.13 (0.89-1.48)	2.07 (1.10-3.78)	< .001
Ferritin, median (IQR), ng/mL	500 (224-1,020)	630	184 (75-426)	567 (295-954)	736 (386-1,305)	372 (196-757)	710 (319-1,568)	< .001
Hematocrit, median (IQR), %	39.6 (35.9-43.3)	118	40.9 (37.3-44.0)	41.8 (39.0-44.2)	40.2 (37.3-43.5)	39.5 (36.2-43.3)	34.6 (29.6-38.9)	< .001
Hemoglobin, median (IQR), g/dL	13.10 (11.50-14.40)	118	13.40 (12.30-14.70)	14.00 (13.00-15.00)	13.40 (12.20-14.60)	12.90 (11.67-14.30)	10.65 (9.20-12.60)	< .001
Platelets, median (IQR), K/μL	207 (161-267)	118	247 (200-318)	192 (162-220)	263 (214-321)	174 (139-218)	197 (138-279)	< .001

(Continued)

TABLE 2] *(Continued)*

Variable	Overall (N = 2,655)	No. of Patients With Missing Data	Subclass 1 (n = 363)	Subclass 2 (n = 568)	Subclass 3 (n = 524)	Subclass 4 (n = 672)	Subclass 5 (n = 5,281)	*P* Value
Potassium, median (IQR), mM	4.00 (3.70-4.30)	144	3.90 (3.60-4.20)	3.90 (3.60-4.20)	3.90 (3.60-4.30)	4.00 (3.70-4.38)	4.30 (3.90-4.70)	< .001
RDW, median (IQR), %	13.80 (12.90-15.00)	365	13.50 (12.70-14.60)	13.20 (12.43-14.00)	13.50 (12.80-14.60)	13.90 (13.20-14.90)	15.20 (13.80-17.00)	< .001
Total bilirubin, median (IQR), mg/dL	0.40 (0.30-0.60)	215	0.40 (0.20-0.60)	0.40 (0.30-0.60)	0.50 (0.40-0.70)	0.40 (0.30-0.60)	0.40 (0.30-0.70)	< .001
WBC count, median (IQR), K/μL	6.5 (4.9-8.7)	118	7.1 (5.0-10.1)	5.3 (4.5-6.5)	8.6 (7.2-10.4)	5.3 (4.3-6.8)	7.7 (5.0-11.4)	< .001
Outcomes								
ICU transfer	903 (34)	0	44 (12)	109 (19)	257 (49)	198 (29)	295 (56)	< .001
In-hospital mortality	249 (9.4)	0	0 (0)	16 (2.8)	53 (10)	69 (10)	111 (21)	< .001
ICU transfer or in-hospital mortality	936 (35)	0	44 (12)	109 (19)	260 (50)	214 (32)	309 (59)	< .001

P values are based on Kruskal-Wallis test, Pearson χ^2 test, or Fisher exact test as appropriate. ACE = angiotensin-converting enzyme; ALT = alanine aminotransferase; ARB = angiotensin receptor blocker; CRP = C-reactive protein; IQR = interquartile range; RDW = RBC distribution width.

albumin, 3.30 g/dL), anemia (hemoglobin, 10.65 g/dL), lymphopenia (median lymphocyte count, 0.88 × 10^9 cells/L), and elevated CRP (median, 9.0 mg/dL). Compared with other subclasses, these patients had the highest likelihood of ICU transfer or in-hospital mortality (59%).

We performed Kaplan-Meier survival analysis for the two time-to-event outcomes by subphenotypes: time to ICU transfer or in-hospital mortality and time to in-hospital mortality only. As shown in Figure 3A, patients in subclasses 3 and 5 had much higher probabilities of being transferred to ICU or dying in hospital than did patients in subclasses 1 and 2. When we investigated the time to in-hospital mortality only, Figure 3B shows that subclasses 4 and 5 had worse overall survival than the other three classes. Patients in subclass 3, although having a high chance of being transferred to the ICU, appeared to have a more favorable survival than subclass 4. The *P* values of the log-rank tests for the two outcomes were < .001 when comparing the five subphenotypes.

Phenotype and Subphenotype Prediction

We established two multivariable prediction models to predict assignment to the LCA-defined phenotypes and subphenotypes, respectively. For phenotypes of patients with COVID-19, we used one demographic variable (age), three symptom variables (cough, fever, and diarrhea), and three comorbidity variables (cancer, diabetes, and hypertension) in the model. With the seven variables, the concordance index (*C* index) for the phenotype prediction, using bootstrapped internal validation, was 0.93.

For subphenotypes of hospitalized patients, we used four of the blood biomarkers with the greatest difference in mean absolute values between phenotypes as predictive markers, and two clinical variables in a predictive modeling analysis. With six variables (creatinine, albumin, CRP, WBC count, age, cancer/COPD/emphysema [yes or no]), the *C* index was 0.91. The estimated parameters of the two prediction models for assignment to the LCA-defined phenotypes and subphenotypes are presented in e-Tables 5 and 6. We have developed an online calculator to predict the LCA-defined phenotypes and subphenotypes (e-Fig 1).[21]

Validation Analysis

We validated the accuracy of outcome prediction, using the phenotype and subphenotype groupings, on

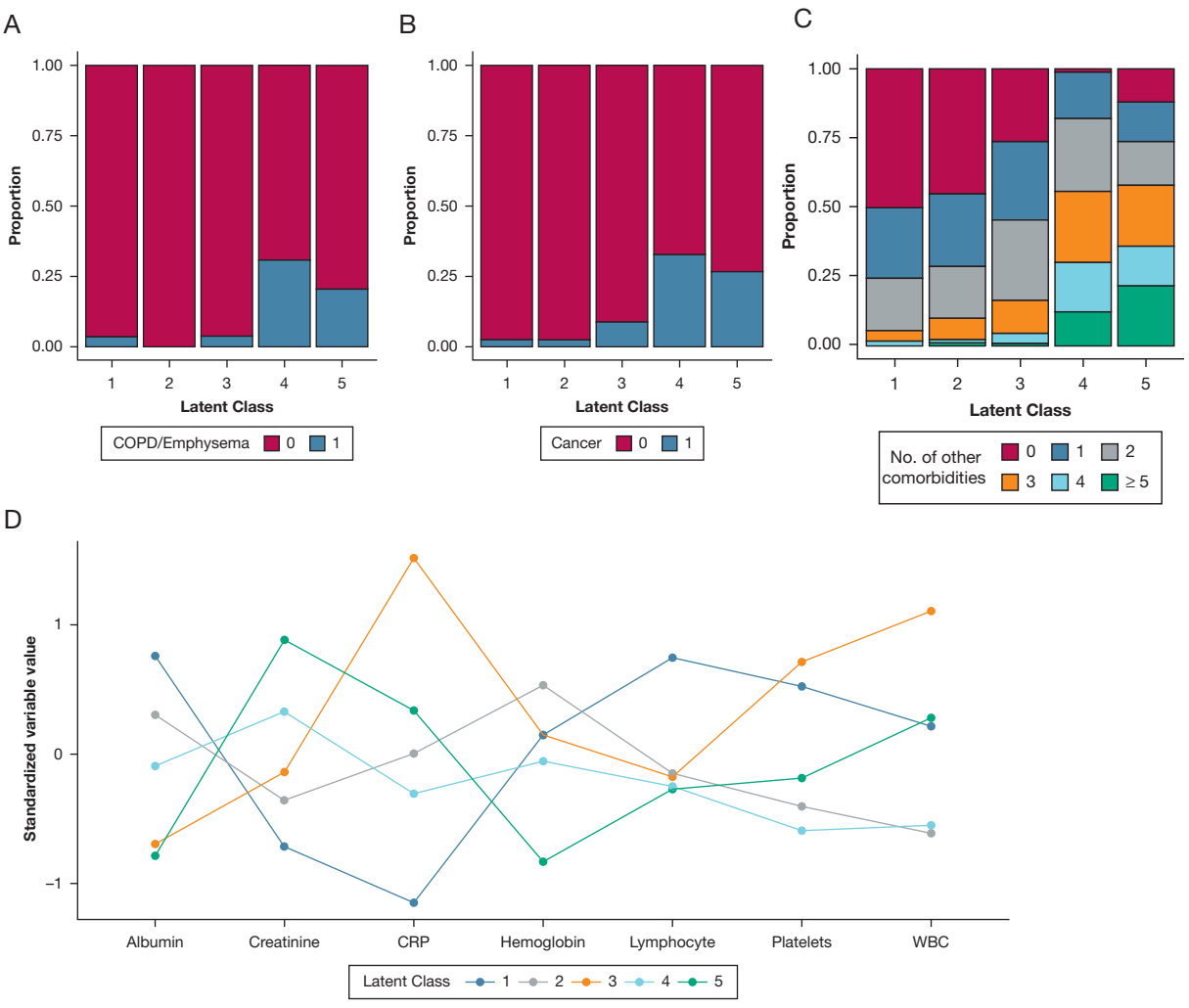

Figure 2 – *Latent profile plots for the class-defining variables for hospitalized patients. CRP = C-reactive protein.*

an independent test set and compared the accuracy with models based on age and the presence of key comorbidities alone. In Table 3, we summarize the *C* indexed of 15 different models applied to the independent testing cohort. The *C* indexes for the models using the phenotype or subphenotype stratification were uniformly better than those of the models using age and/or the presence of key comorbidities.

Discussion

This study, the first to include a large cohort of patients with COVID-19 with comprehensive demographic, clinical, and blood biomarker data, exclusively focused on objective disease subclassification into clinical phenotypes. We applied latent class modeling, a modern statistical technique that allows us to capture the clinical and biological heterogeneity of patients with COVID-19. Our findings provide proof-of-concept that the clinical spectrum of COVID-19 contains distinct subphenotypes based on patients' clinical information and blood biomarkers available on hospital admission.

Prior publications have identified isolated predictors of severe disease progression, but such information has limited potential to impact clinical decision-making significantly. In fact, no single clinical variable or biomarker was sufficient to identify the subphenotypes in our study, which translates into the reality of clinical practice: patients do not usually fall into clean buckets of "all good" or "all bad" outcome predictors, and clinical outcomes are rarely—if ever—driven by one patient characteristic. However, when considered together, the variables identified in our study form plausible and coherent clusters of patients with COVID-19, and our LCA-defined phenotypes and subphenotypes may indicate different disease mechanisms.

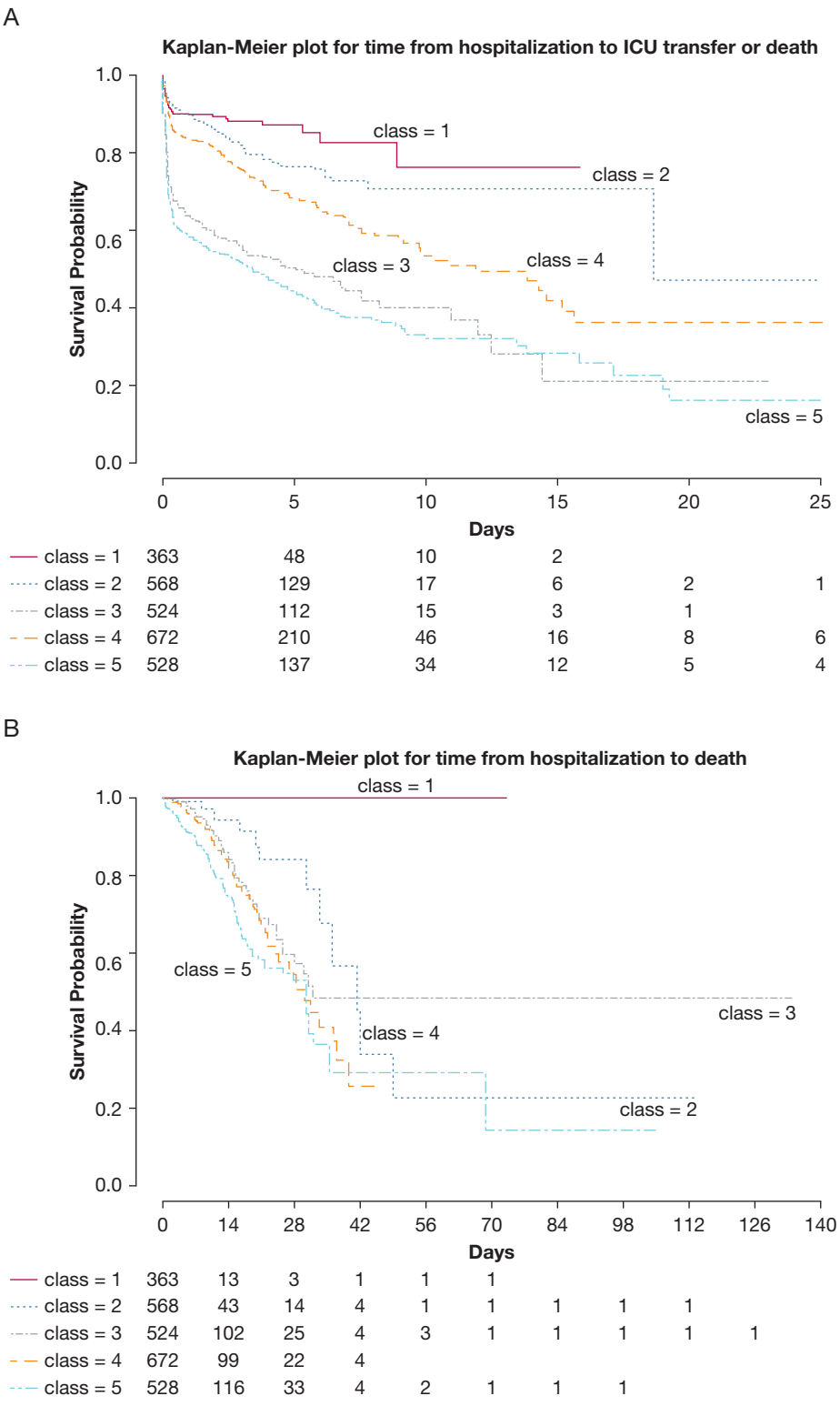

Figure 3 – *Kaplan-Meier plots for two time-to-event outcomes for hospitalized patients. The* P *values of log-rank tests were* < *.001 in comparisons of the five subphenotypes. A, Kaplan-Meier plot for time from hospitalization to ICU transfer or death. B, Kaplan-Meier plot for time from hospitalization to death.*

Our findings provide an objective phenotypic classification reflecting the emerging literature on critical drivers of disease pathophysiology in COVID-19. In summary, of the phenotypes of patients who tested positive, classes 1 and 2 included young people with or without symptoms: about 10% were hospitalized. Classes

TABLE 3] Comparison of Prediction Performance, Using Phenotype or Subphenotype vs Age and/or Presence of Cancer/COPD/Emphysema in Prediction Models: Fifteen Different Models Applied to Testing Cohort in Predicting Different Outcomes

Outcome	Covariate(s)	*C* Index	Covariate(s)	*C* Index	Covariate(s)	*C* Index
Hospitalization	Phenotype	0.77	Age group	0.70	Age	0.70
Hospitalization	Phenotype + cancer/COPD/emphysema	0.78	Age group + cancer/COPD/emphysema	0.73	Age + cancer/COPD/emphysema	0.73
Time from hospitalization to ICU transfer or in-hospital mortality	Subphenotype	0.63	Age group	0.56	Age	0.55
Time from hospitalization to ICU transfer or in-hospital mortality	Subphenotype + cancer/COPD/emphysema	0.64	Age group + cancer/COPD/emphysema	0.57	Age + cancer/COPD/emphysema	0.55
Time from hospitalization to ICU transfer or in-hospital mortality	Subphenotype + cancer/COPD/emphysema + ALT	0.65	Age group + cancer/COPD/emphysema + ALT	0.59	Age + cancer/COPD/emphysema + ALT	0.58

"Age group" denotes the categorical variable that patient's age was categorized by decade; "Age" denotes the continuous variable of patient's age; "ALT" is in a logarithmic scale in the model. ALT = alanine aminotransferase; *C* index = concordance index.

3, 4, and 5 consisted of middle-aged people with or without symptoms and different comorbidity patterns: 22% to 33% were hospitalized. Classes 6 and 7 included old adults: if they had no or few symptoms, the hospitalization rate was close to 50%. If they had numerous symptoms, the rate was close to 60%.

For subphenotypes of hospitalized patients, the findings highlight two points: the first is the importance of the individual immune response in stratifying disease presentation and driving outcomes. As recently published by Zeng et al,[22] the levels of proinflammatory cytokines, lactate dehydrogenase, and CRP are higher within 24 h of hospitalization and do not recover over the subsequent 10 days in those who are critically ill or die when compared with their counterparts with moderate or severe disease. Similar observations have been published elsewhere,[23,24] emphasizing that the individual patient immune phenotype heavily drives morbidity and mortality with COVID-19, and not just cellular damage directly inflicted by the virus. Our findings in this article provide a tool to put this immune response (as reflected by admission laboratory test results) into clinical context with other patient characteristics. Second, kidney function emerged as another important clinical classifier and outcome determinant, again consistent with published literature findings that renal failure on admission in patients with SARS-CoV-2 infection is frequent and associated with a greater number of complications and in-hospital mortality.[25] Table 3 demonstrates that our LCA phenotypic model is a better predictor of hospitalization and of progression to ICU admission or mortality than a risk assessment based on age and comorbidities alone. This supports an incremental benefit to our model over existing knowledge.

In addition, our findings in Figures 3A and 3B suggest potentially relevant clinical implications. For example, subclass 2 has the second-highest survival probability in Figure 3A but tied for the lowest in Figure 3B, raising the hypothesis that late transfer to the ICU may have influenced/delayed the death outcome in this class. It is possible that late complications (thromboembolic events or secondary infections) were potentially avoided by higher care intensity in the ICU. Similar questions could apply to subclass 4. Conversely, members of subclass 5 were quickest to get to the ICU or die (Fig 3A) and had the highest death rate (Fig 3B). Is ICU transfer or transfer time futile for this group? Individuals in subclass 3 also had a fast transfer time but a substantially lower rate of death than subclasses 5, 4, and 2. Should this be a group we focus our resources on? These are all speculations at this point, but warrant further investigations with future studies.

Our study has some limitations. First, we did not include radiologic features in the LCA. Robba et al[26] described three distinct radiologic phenotypes of severe COVID-19 pneumonia and suggested that their radiologic

phenotypes could redefine clinical management. Radiologic features of COVID-19 may add to the power to improve classification performance[27] and the precision of parameter estimates in the LCA. However, radiologic analysis may be delayed compared with the variables included, which would decrease the impact of using the latent class models. Second, our analyses of possible classifying variables were restricted to the data obtained in the study. We did not have complete information regarding the need for oxygen support, cardiac involvement, and thromboembolic disease. We did not include biomarker variables with a high proportion of missing values, such as lactate and procalcitonin. Overweight and obese patients are at risk for many medical conditions, which can lead to further morbidity and mortality, which might be associated with the prognosis of COVID-19. We were unable to consider them in our analysis because BMI data were missing for 47% of the cohort. These variables could contribute to phenotype identification but were not available for this analysis. Finally, we considered that our COVID-19 data registry is a valid source of data for the assessment of comorbidities at a population-based level. Given the chronic aspect of comorbidities, we assumed that the patient has that comorbidity up until the time of COVID-19 diagnosis once a comorbidity is recorded in our electronic health record. However, the evaluation of comorbidities depends heavily on both age and the probability of attending the hospital (inpatient or outpatient) in the years preceding the COVID-19 diagnosis. We did not systematically evaluate the accuracy of data about comorbidities and smoking history, and did not have data about the severity of the comorbid conditions. A limitation of the outcome used to assess the first latent class model is that the decision about who to hospitalize is not standardized. We also lacked laboratory test results over time, which limited us to study the disease progression.

At present, the immunogenicity, efficacy, safety, and production capacity of COVID-19 vaccines are not yet clear. Many clinical research studies are still on-going. The responses to treatment by the phenotypes or subphenotypes could vary; thus our results can inform future randomized controlled trials of novel therapies for COVID-19. The identification of the phenotypes has the potential to help investigators draw more robust conclusions about causal associations between treatment and outcomes in clinical trials through phenotype group analyses, as well as assist basic science researchers in characterizing the pathobiology of COVID-19. Future research could be done to further understand the underlying mechanisms of our phenotypic subclassification (ideally through host and virus genomics) and explore the treatment implications.

Acknowledgments

Author contributions: X. W. and P. J. M. had the idea for and designed the study. X. W., L. J., and P. J. M. drafted the article. X. W. and X. J. had full access to all of the data in the study, take responsibility for the integrity of the data, and performed the data analysis. All of the authors critically revised the manuscript for important intellectual content and gave final approval for the version to be published. All of the authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Financial/nonfinancial disclosures: None declared.

Additional information: The e-Figure and e-Tables can be found in the Supplemental Materials section of the online article.

References

1. Grasselli G, Zangrillo A, Zanella A, et al. COVID-19 Lombardy ICU Network. Baseline characteristics and outcomes of 1591 patients infected with SARS-CoV-2 admitted to ICUs of the Lombardy region, Italy. *JAMA*. 2020;323(16):1574-1581.
2. Richardson S, Hirsch JS, Narasimhan M, et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. *JAMA*. 2020;323(20):2052-2059.
3. Wang D, Hu B, Hu C, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA*. 2020;323(11):1061-1069.
4. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet*. 2020;395(10229):1054-1062.
5. Guan W-J, Ni Z-Y, Hu Y, et al; China Medical Treatment Expert Group for Covid-19. Clinical characteristics of coronavirus disease 2019 in China. *N Engl J Med*. 2020;382(18):1708-1720.
6. Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72314 cases from the Chinese Center for Disease Control and Prevention. *JAMA*. 2020;323(13):1239-1242.
7. Rello J, Storti E, Belliato M, Serrano R. Clinical phenotypes of SARS-CoV-2: implications for clinicians and researchers. *Eur Respir J*. 2020;55(5):2001028.
8. Gattinoni L, Chiumello D, Caironi P, et al. COVID-19 pneumonia: different respiratory treatments for different phenotypes? *Intensive Care Med*. 2020;46(6):1099-1102.
9. Depner M, Fuchs O, Genuneit J, et al. Clinical and epidemiologic phenotypes of childhood asthma. *Am J Respir Crit Care Med*. 2014;189(2):129-138.
10. Sakr L, Small D, Kasymjanova G, Suissa S, Ernst P. Phenotypic heterogeneity of potentially curable non-small-cell lung cancer: cohort study with cluster analysis. *J Thorac Oncol*. 2015;10(5):754-761.
11. Calfee CS, Delucchi K, Parsons PE, et al. Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. *Lancet Respir Med*. 2014;2(8):611-620.
12. Hagenaars JA, McCutcheon AL. *Applied Latent Class Analysis*. Cambridge: Cambridge University Press; 2002.
13. Jehi L, Ji X, Milinovich A, et al. Individualizing risk prediction for positive COVID-19 testing: results from 11,672 patients. *Chest*. 2020;158(4):1364-1375.
14. Agency for Healthcare Research and Quality (AHRQ). AHRQ Quality Indicators: Patient Safety Indicators (PSI) Appendix I: Immunocompromised State Diagnosis and Procedure Codes. https://qualityindicators.ahrq.gov/ICD10/default.aspx. Accessed February 5, 2021.
15. Zhang L, Zhu F, Xie L, et al. Clinical characteristics of COVID-19-infected cancer patients: a retrospective case study in three hospitals within Wuhan, China. *Ann Oncol*. 2020;31(7):894-901.
16. Lippi G, Henry BM. Chronic obstructive pulmonary disease is associated with severe coronavirus disease 2019 (COVID-19). *Respir Med*. 2020;167:105941.
17. Fop M, Smart KM, Murphy TB. Variable selection for latent class analysis with application to low back pain diagnosis. *Ann Appl Stat*. 2017;11(4):2080-2110.
18. Visser I, Speekenbrink M. depmixS4: an R package for hidden Markov models. *J Stat Softw*. 2010;36(7):1-21.
19. Bolck A, Croon M, Hagenaars J. Estimating latent structure models with categorical variables: one-step versus three-step estimators. *Pol Anal*. 2004;12(1):3-27.
20. Vermunt JK. Latent class modeling with covariates: two improved three-step approaches. *Pol Anal*. 2010;18(4):450-469.
21. Department of Quantitative Health Sciences, Cleveland Clinic Risk Calculator Library. Online calculator for phenotype prediction for patients with COVID-19. https://riskcalc.org/COVID19phenotypes/. Accessed February 18, 2021.
22. Zeng Z, Yu H, Chen H, et al. Longitudinal changes of inflammatory parameters and their correlation with disease severity and outcomes in patients with COVID-19 from Wuhan, China. *Crit Care*. 2020;24(1):1-12.
23. Xu X, Yu M-Q, Shen Q, et al. Analysis of inflammatory parameters and disease severity for 88 hospitalized COVID-19 patients in Wuhan, China. *Int J Med Sci*. 2020;17(13):2052.
24. Wu H, Zhu H, Yuan C, et al. Clinical and immune features of hospitalized pediatric patients with coronavirus disease 2019 (COVID-19) in Wuhan, China. *JAMA Netw Open*. 2020;3(6):e2010895.
25. Uribarri A, Núñez-Gil IJ, Aparisi A, et al. Impact of renal function on admission in COVID-19 patients: an analysis of the international HOPE COVID-19 (Health Outcome Predictive Evaluation for COVID 19) Registry. *J Nephrol*. 2020;33(4):737-745.
26. Robba C, Battaglini D, Ball L, et al. Distinct phenotypes require distinct respiratory management strategies in severe COVID-19. *Respir Physiol Neurobiol*. 2020;279:103455.
27. Wang X, Hu X, Tan W, et al. Multi-center study of temporal changes and prognostic value of a CT visual severity score in hospitalized patients with COVID-19. *AJR Am J Roentgenol*. In press.