npj | Digital Medicine www.nature.com/npjdigitalmed

ARTICLE OPEN

Clinical subphenotypes in COVID-19: derivation, validation, prediction, temporal patterns, and interaction with social determinants of health

Chang Su[1], Yongkang Zhang[1], James H. Flory[2], Mark G. Weiner[1], Rainu Kaushal[1,3,4]✉, Edward J. Schenck[3,5]✉ and Fei Wang[1]✉

The coronavirus disease 2019 (COVID-19) is heterogeneous and our understanding of the biological mechanisms of host response to the viral infection remains limited. Identification of meaningful clinical subphenotypes may benefit pathophysiological study, clinical practice, and clinical trials. Here, our aim was to derive and validate COVID-19 subphenotypes using machine learning and routinely collected clinical data, assess temporal patterns of these subphenotypes during the pandemic course, and examine their interaction with social determinants of health (SDoH). We retrospectively analyzed 14418 COVID-19 patients in five major medical centers in New York City (NYC), between March 1 and June 12, 2020. Using clustering analysis, 4 biologically distinct subphenotypes were derived in the development cohort ($N = 8199$). Importantly, the identified subphenotypes were highly predictive of clinical outcomes (especially 60-day mortality). Sensitivity analyses in the development cohort, and rederivation and prediction in the internal ($N = 3519$) and external ($N = 3519$) validation cohorts confirmed the reproducibility and usability of the subphenotypes. Further analyses showed varying subphenotype prevalence across the peak of the outbreak in NYC. We also found that SDoH specifically influenced mortality outcome in Subphenotype IV, which is associated with older age, worse clinical manifestation, and high comorbidity burden. Our findings may lead to a better understanding of how COVID-19 causes disease in different populations and potentially benefit clinical trial development. The temporal patterns and SDoH implications of the subphenotypes may add insights to health policy to reduce social disparity in the pandemic.

npj Digital Medicine (2021)4:110 ; https://doi.org/10.1038/s41746-021-00481-w

INTRODUCTION

The outbreak of coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection, has led to a pandemic that imposed tremendous pressure on healthcare systems globally[1]. As the pandemic continues and the second wave has emerged in the US and many other countries, research is still needed to understand how SARS-CoV-2 causes the wide spectrum of COVID-19 disease. Previous studies have uncovered substantial variation in the host response to SARS-CoV-2 and the variable clinical manifestations of this disease, including respiratory failure, kidney injury, and cardiovascular dysfunction[2–8]. Pivotal studies of corticosteroids[9] and anticoagulation[10,11] demonstrate differential responses in distinct subpopulations based on severity of disease. The pathophysiology of differential organ dysfunction in COVID-19 remains unclear across varied patient populations. Prior to the COVID-19 pandemic, identification of biologically distinct, data-driven subphenotypes[12,13] has helped to disentangle complex syndromic disease, such as sepsis[14,15], ARDS[16], heart failure[17,18], diabetes[19], and Alzheimer's disease[20].

Identifying robust subphenotypes in COVID-19 patients could lead to improved understanding of biological mechanisms of host response to SARS-CoV-2 infection and may identify subpopulations that could be prioritized for clinical trial enrollment[13,21]. Previous efforts[22–25] have been made in this area but remain limited probably due to cohort size, data availability, and lacking evaluation of robustness and usability of the identified subphenotypes. In addition, the hospitalized case fatality rate of COVID-19 has varied over the course of the pandemic[26,27] and according to social determinants of health (SDoH)[28–30]. Exploration of temporal patterns and SDoH characteristics in conjunction with subphenotypes may derive new insights to improve public health.

In this analysis, our goal was to derive and validate COVID-19 subphenotypes amongst a population of patients who presented to the emergency department (ED) or were hospitalized in multiple health systems in New York City (NYC). Specifically, we used routinely collected clinical data to first derive subphenotypes using the agglomerative hierarchical clustering model. Then, multiple strategies in data preprocessing, data filtering, and data-driven models (both unsupervised clustering model and supervised predictive model) were used to confirm reproducibility and usability of the identified subphenotypes. After that, statistical analyses were conducted to evaluate the characteristics and clinical outcomes of the subphenotypes. Further analyses were performed to examine temporal patterns of the subphenotypes and impacts of SDoH status on subphenotype-level outcomes. The overall workflow of our study is illustrated in Fig. 1.

[1]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA. [2]Memorial Sloan-Kettering Cancer Center, New York, NY, USA. [3]New York-Presbyterian Hospital, Weill Cornell Medicine, New York, NY, USA. [4]Department of Medicine, Weill Cornell Medical College, New York, NY, USA. [5]Division of Pulmonary & Critical Care Medicine, Joan and Sanford I. Weill Department of Medicine, Weill Cornell Medicine, New York, NY, USA. ✉email: rak2007@med.cornell.edu; ejs9005@med.cornell.edu; few2001@med.cornell.edu

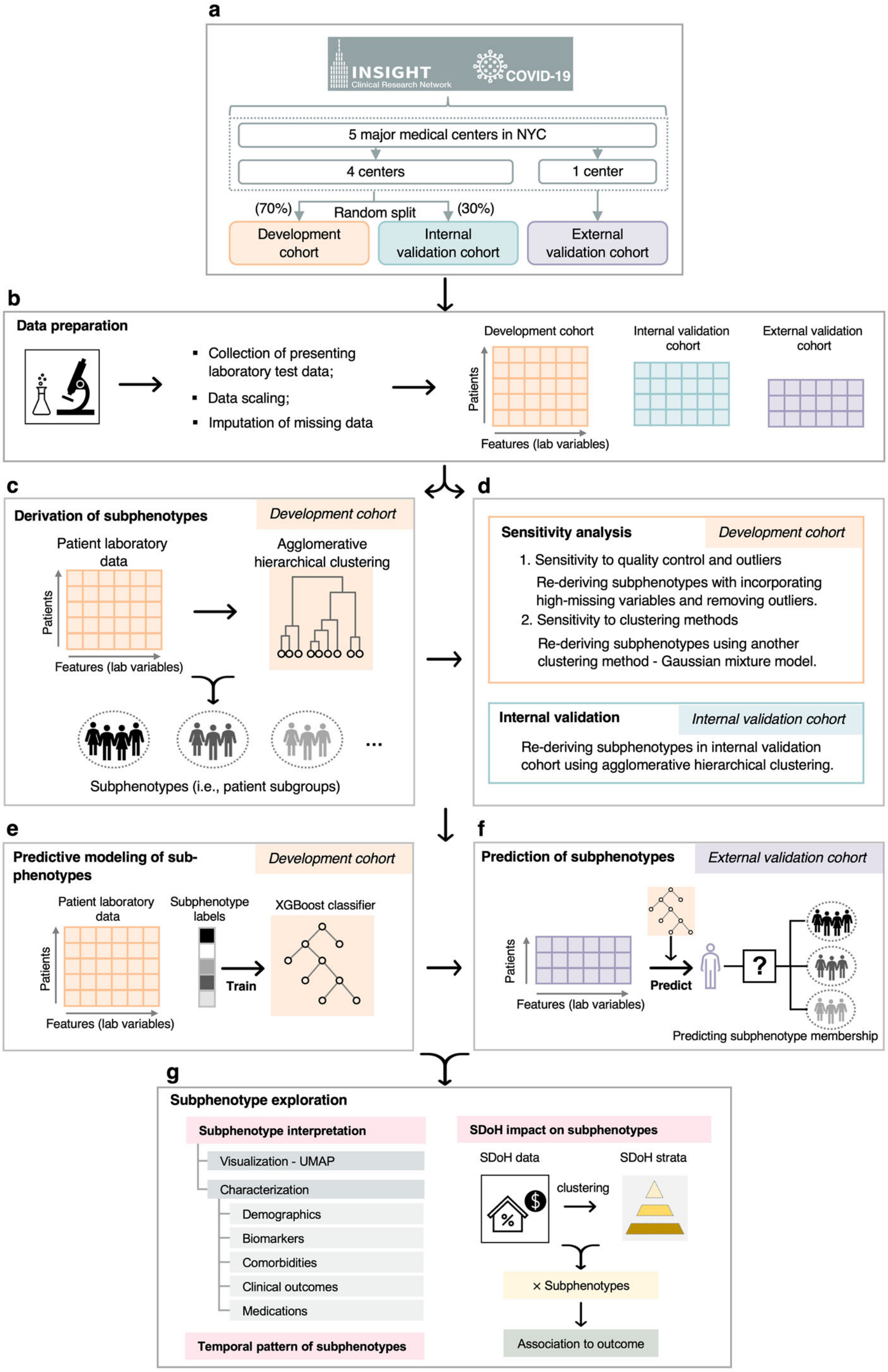
a
INSIGHT Clinical Research Network
COVID-19
5 major medical centers in NYC
4 centers
1 center
(70%)
Random split
(30%)
Development cohort
Internal validation cohort
External validation cohort
b
Data preparation
Collection of presenting laboratory test data;
Data scaling;
Imputation of missing data
Development cohort
Internal validation cohort
External validation cohort
Patients
Features (lab variables)
c
Derivation of subphenotypes
Development cohort
Patient laboratory data
Agglomerative hierarchical clustering
Patients
Features (lab variables)
...
Subphenotypes (i.e., patient subgroups)
d
Sensitivity analysis
Development cohort
1. Sensitivity to quality control and outliers
Re-deriving subphenotypes with incorporating high-missing variables and removing outliers.
2. Sensitivity to clustering methods
Re-deriving subphenotypes using another clustering method - Gaussian mixture model.
Internal validation
Internal validation cohort
Re-deriving subphenotypes in internal validation cohort using agglomerative hierarchical clustering.
e
Predictive modeling of sub-phenotypes
Development cohort
Patient laboratory data
Subphenotype labels
XGBoost classifier
Patients
Train
Features (lab variables)
f
Prediction of subphenotypes
External validation cohort
Patients
Predict
?
Features (lab variables)
Predicting subphenotype membership
g
Subphenotype exploration
Subphenotype interpretation
Visualization - UMAP
Characterization
Demographics
Biomarkers
Comorbidities
Clinical outcomes
Medications
Temporal pattern of subphenotypes
SDoH impact on subphenotypes
SDoH data
SDoH strata
clustering
%
$
× Subphenotypes
Association to outcome

Fig. 1 A schematic of the analysis plan. **a** Strategy for construction of development, internal validation, and external validation cohorts. **b** Data preparation for clustering analysis. **c** Derivation of subphenotypes in the development cohort. Reproducibility of the identified subphenotypes were evaluated in multiple ways, including **d** sensitivity analyses in the development cohort and subphenotype rederivation in the internal validation cohort; and **e** training subphenotype predictive model in the development cohort and **f** using it to predict subphenotype memberships of patients in the external validation cohort. Last, **g** further analyses were conducted to interpret subphenotypes, explore temporal patterns of subphenotypes during the pandemic, and evaluate impact of SDoH characterisitics on subphenotypes. NYC New York City, SDoH social determinants of health, UMAP Uniform Manifold Approximation and Projection.

RESULTS

Patients

A total of 14418 patients with confirmed COVID-19 between March 1 and June 12, 2020, treated in ED ($N = 2354$, 16.3%) or inpatient ($N = 12064$, 83.7%) settings, were included for analysis from the five major medical centers in New York City (NYC). Details of inclusion and exclusion criteria are presented in the Supplementary Fig. 1. We identified 2853 (19.8%) deaths within 60-day after COVID-19 confirmation in total, including 2801 (19.4%) in-hospital deaths and 52 (4%) deaths after discharge from COVID related hospitalization or ED visits. Considering population diversity (especially race) of the five medical centers (see Supplementary Table 1), we combined four centers and randomly divided them into the development cohort (70%) and

Table 1. Characteristics of the development, internal validation, and external validation cohorts.

Characteristics	Cohort		
	Development cohort	Internal validation cohort	External validation cohort
No. of patients	8199	3519	2700
Construction method	70% patients (randomly selected) from 4 medical centers	Remaining 30% patients from 4 medical centers	Patients from the last center
Age, y, Median (IQR)	63.53 [50.57–75.15]	63.51 [50.95–75.17]	65.58 (51.08–77.39)
Sex female, *N* (%)	3787 (46.2)	1585 (45.0)	1305 (48.3)
Race, *N* (%)			
White	2036 (24.8)	838 (23.8)	675 (25.0)
Black	2155 (26.3)	915 (26.0)	545 (20.2)
Asian	409 (5.0)	193 (5.5)	28 (1.0)
Other/unknown	3599 (43.9)	1573 (44.7)	1452 (53.8)
Outcomes (60 days), *N* (%)			
Mortality	1529 (18.65)	696 (19.78)	556 (20.59)
Mechanical ventilation (intubation	1154 (14.07)	497 (14.12)	248 (9.19)
ICU admission	1494 (18.22)	661 (18.78)	–

ICU intensive care unit, *IQR* interquartile range, *SDoH* social determinants of health.

internal validation cohort (30%); patients of the remaining center were used as the external validation cohort (see Fig. 1 and Supplementary Fig. 1).

The development cohort contained a total of 8199 patients with a median age of 65.35 (interquartile range [IQR] [50.57, 75.17]) years old, consisting of 3787 (46.2%) females, 2036 (24.8%) white patients, and 2155 (26.3%) black patients. The internal validation cohort contained a total of 3519 patients with similar patient characteristics when compared with the development cohort, with a median age of 63.51 (IQR [50.95, 75,17]) years old, consisting of 1585 (45.0%) females, 838 (23.8%) white patients, and 915 (26%) black patients. The external validation cohort contained a total of 2700 patients. It had a median age of 65.85 (IQR [51.08, 77.38]) years old and consisted of 1305 (48.3%) females, 675 (25.0%) white patients, and 545 (20.2%) black patients. Across the three cohorts, the overall 60-day mortality rates after ED or hospital discharge were 18.65, 19.78, and 20.59%, respectively. More details of the characteristics of the studied cohorts appeared in Table 1.

Subphenotypes derivation

In the development cohort, the agglomerative hierarchical clustering model identified 4 distinct subphenotypes based on presenting clinical data of the patients (see Supplementary Figs. 3 and 4 and Supplementary Results). Characteristics including demographics, clinical variables, comorbidities, clinical outcomes, and medication treatments across the 4 subphenotypes were presented in Table 2 and Figs. 2 and 3.

Subphenotype I consisted of 2707 (33.02%) patients. Compared to the others, it included more younger (median age 57.45 years, IQR [42.70, 70.02]) and female (*N* = 1601 [59.15%]) patients. Those patients had more normal values across all clinical variables and lower chronic comorbidity burden. The patients also had better clinical outcomes including a low 60-day mortality (*N* = 188 [6.94%]) and a low rates of mechanical ventilation (*N* = 190 [7.02%]) and ICU admission (*N* = 242 [8.94%]).

Subphenotype II consisted of 3047 (37.16%) patients. Compared to other subphenotypes, it included more male patients (*N* = 1941 [63.70%]) and was likely to have more abnormal inflammatory markers (such as C-reactive protein, erythrocyte sedimentation rate, interleukin 6, lactate dehydrogenase, lymphocyte count, neutrophil count, white blood cell count, and ferritin), and markers of hepatic dysfunctions (such as ferritin, alanine aminotransferase, aspartate aminotransferase, and bilirubin). Overall comorbidity burden of Subphenotype II was low. Clinical outcomes including 60-day mortality (*N* = 528 [17.33%]), mechanical ventilation (*N* = 527 [17.30%]), and ICU admission (*N* = 675 [22.15%]) of Subphenotype II were at a moderate level.

Subphenotype III included 1486 (18.12%) patients, consisting of more older (median age 69.45 years, IQR [57.05, 79.62]) and black (*N* = 503 [33.85%]) patients, compared to subphenotypes I and II. Those patients of Subphenotype III were likely to have more abnormal renal dysfunction markers (such as blood urea nitrogen, creatinine, chloride, and sodium) and hematologic dysfunction markers (such as d-dimer, hemoglobin, and red blood cell distribution width). Overall comorbidity burden of Subphenotype III was high. Clinical outcomes including 60-day mortality (*N* = 337 [22.68%]), intubation (*N* = 195 [13.12%]), and ICU admission (*N* = 242 [16.29%]) of Subphenotype II were close to that of Subphenotype II and at a moderate level as well.

Subphenotype IV included 959 (11.70%) patients. Compared to other subphenotypes, it included more older (median age 75.53 years, IQR [64.10, 84.83]) and male (*N* = 588 [61.31%]) patients. Those patients of Subphenotype IV had more abnormal values across all clinical variables and higher chronic comorbidity burden than the others. Obesity burden is lower in Subphenotype IV than others. In line with its biological characteristics, Subphenotype IV had the worst clinical outcomes in 60-day mortality (*N* = 476 [49.64%]), intubation (*N* = 242 [25.23%]), and ICU admission (*N* = 335 [34.93%]). In addition, the medications including antibiotics, corticosteroids, and vasopressor were more frequently used in Subphenotype IV.

Subphenotype reproducibility and prediction

In the development cohort, sensitivity analyses under two different settings (sensitivity to quality control and outliers and sensitivity to clustering methods) confirmed the underlying 4-cluster structure of the data (see Supplementary Figs. 3 and 4, Supplementary Table 5, and Supplementary Results). Patients' memberships of the 4 clusters rederived by sensitivity analyses were highly consistent with those derived in the primary analysis (see Supplementary Fig. 6). Moreover, we did not find substantial

Table 2. Characteristics of the identified subphenotypes (development cohort).

Variable	Total	Subphenotype I	Subphenotype II	Subphenotype III	Subphenotype IV	*P*-value[a]	*P*-value (age and sex adjusted)[b]
No. of patients (%)	8199 (100)	2707 (33.02)	3047 (37.16)	1486 (18.12)	959 (11.70)	–	–
Demographics							
Age, y, Median (IQR)	63.53 (50.57–75.15)	57.45 (42.70–70.02)	62.56 (51.63–72.77)	69.45 (57.05–79.62)	73.53 (64.10–82.83)	<0.001	–
Sex female, *N* (%)	3787 (46.19)	1601 (59.14)	1106 (36.30)	709 (47.71)	371 (38.69)	<0.001	–
Race, *N* (%)							
White	2036 (24.83)	695 (25.67)	777 (25.50)	367 (24.70)	197 (20.54)	<0.001	–
Black	2155 (26.28)	697 (25.75)	611 (20.05)	503 (33.85)	344 (35.87)		
Asian	409 (4.99)	118 (4.36)	194 (6.37)	58 (3.90)	39 (4.07)		
Other/unknown	3599 (43.90)	1197 (44.22)	1465 (48.08)	558 (37.55)	379 (39.52)		
Inflammatory markers							
C-reactive protein, mg/L, Median (IQR)	9.40 (3.70–16.80)	4.32 (1.16–9.31)	12.74 (6.60–20.20)	8.20 (3.50–14.51)	14.90 (6.70–23.07)	<0.001	<0.001
ESR, mm/h, Median (IQR)	69.00 (42.00–97.00)	53.00 (34.00–81.00)	76.00 (50.00–100.00)	75.00 (45.25–102.75)	77.00 (41.75–106.25)	<0.001	<0.001
IL-6, pg/mL, Median (IQR)	19.00 (10.00–42.00)	13.00 (8.00–21.00)	21.00 (11.00–45.75)	17.00 (9.00–47.00)	27.00 (10.25–52.00)	<0.001	0.26
Procalcitonin, ng/mL, Median (IQR)	0.20 (0.10–0.60)	0.10 (0.10–0.20)	0.20 (0.10–0.50)	0.30 (0.10–0.87)	0.60 (0.25–2.10)	<0.001	0.04
Bands, %, Median (IQR)	2.00 (0.00–5.00)	3.00 (0.00–5.75)	2.00 (0.00–5.00)	2.00 (0.00–5.00)	2.00 (0.00–6.00)	0.37	0.14
LDH, U/L, Median (IQR)	377.00 (280.00–525.00)	292.00 (229.00–377.00)	437.00 (343.00–576.00)	349.00 (268.00–449.00)	565.50 (409.75–801.50)	<0.001	<0.001
Lymphocyte count, $\times 10^3$/uL, Median (IQR)	1.00 (0.70–1.43)	1.20 (0.80–1.60)	1.00 (0.70–1.40)	0.80 (0.60–1.20)	0.90 (0.60–1.40)	<0.001	0.02
Neutrophil count, $\times 10^3$/uL, Median (IQR)	5.30 (3.70–7.90)	4.00 (2.90–5.40)	6.70 (4.80–9.50)	4.70 (3.40–6.60)	8.20 (5.90–11.00)	<0.001	<0.001
White blood cell count, $\times 10^3$/uL, Median (IQR)	7.20 (5.30–9.90)	5.90 (4.60–7.60)	8.50 (6.50–11.50)	6.30 (4.70–8.30)	10.30 (7.60–13.57)	<0.001	<0.001
Inflammation and hepatic markers							
Albumin, g/dL, Median (IQR)	3.70 (3.30–4.10)	4.00 (3.60–4.30)	3.70 (3.20–4.00)	3.50 (3.10–3.90)	3.40 (2.90–3.80)	<0.001	<0.001
Ferritin, ng/mL, Median (IQR)	645.00 (295.90–1347.00)	323.05 (157.75–594.33)	868.80 (454.00–1537.50)	599.00 (217.80–1380.50)	1174.00 (523.00–2284.00)	<0.001	<0.001
Hepatic markers							
Alanine aminotransferase, U/L, Median (IQR)	29.00 (19.00–48.00)	24.00 (17.00–36.00)	41.00 (26.00–68.00)	20.00 (13.00–29.00)	37.00 (22.00–65.00)	<0.001	<0.001
Aspartate aminotransferase, U/L, Median (IQR)	39.00 (26.00–63.00)	31.00 (23.00–42.00)	52.00 (35.00–80.00)	31.00 (22.00–46.00)	65.00 (36.00–118.00)	<0.001	<0.001
Bilirubin, mg/dL, Median (IQR)	0.30 (0.20–0.60)	0.20 (0.20–0.40)	0.40 (0.20–0.70)	0.30 (0.20–0.50)	0.40 (0.20–0.70)	<0.001	<0.001
Cardiovascular markers						<0.001	<0.001
Creatine kinase, U/L, Median (IQR)	154.00 (78.00–359.00)	122.00 (72.00–227.00)	165.00 (83.00–387.50)	126.00 (63.00–288.00)	352.00 (137.00–1039.50)	<0.001	<0.001
Lactate, mmol/L, Median (IQR)	1.90 (1.40–2.60)	1.50 (1.20–2.10)	2.00 (1.50–2.70)	1.60 (1.20–2.10)	3.10 (2.20–4.80)	<0.001	<0.001
Troponin I, ng/mL, Median (IQR)	0.10 (0.06–0.30)	0.10 (0.00–0.10)	0.10 (0.06–0.30)	0.10 (0.10–0.21)	0.20 (0.10–0.50)	<0.001	0.16
Troponin T, ng/mL, Median (IQR)	0.01 (0.01–0.03)	0.01 (0.01–0.01)	0.01 (0.01–0.01)	0.03 (0.01–0.09)	0.05 (0.01–0.14)	<0.001	<0.001

Table 2 continued

Variable	Total	Subphenotype I	Subphenotype II	Subphenotype III	Subphenotype IV	*P*-value[a]	*P*-value (age and sex adjusted)[b]
Renal markers							
Bicarbonate, mmol/L, Median (IQR)	23.00 (21.00–26.00)	24.00 (22.00–27.00)	23.00 (21.00–25.00)	23.00 (20.00–25.00)	20.00 (17.00–23.00)	<0.001	<0.001
BUN, mg/dL, Median (IQR)	17.00 (11.00–31.00)	12.00 (9.00–17.00)	16.00 (12.00–24.00)	31.00 (18.00–53.00)	52.00 (32.00–84.00)	<0.001	<0.001
Creatinine, mg/dL, Median (IQR)	1.00 (0.80–1.50)	0.86 (0.70–1.04)	1.00 (0.80–1.29)	1.70 (1.00–4.40)	2.10 (1.38–3.60)	<0.001	<0.001
Chloride, mmol/L, Median (IQR)	100.00 (97.00–104.00)	101.00 (98.00–104.00)	99.00 (95.00–102.00)	101.00 (97.00–105.00)	104.00 (98.00–113.00)	<0.001	<0.001
Sodium, mmol/L, Median (IQR)	137.00 (134.00–140.00)	138.00 (136.00–140.00)	136.00 (132.00–138.00)	138.00 (134.00–141.00)	141.00 (136.00–152.00)	<0.001	<0.001
Hematologic markers							
D-dimer, ng/mL, Median (IQR)	1360.00 (620.00–3370.00)	660.00 (370.00–1310.00)	1390.00 (690.00–3210.00)	1740.00 (836.50–3520.00)	4000.00 (2000.00–13582.50)	<0.001	<0.001
Hemoglobin, g/dL, Median (IQR)	13.10 (11.50–14.60)	13.40 (12.30–14.60)	13.80 (12.50–15.10)	10.80 (9.00–12.30)	12.75 (10.70–15.00)	<0.001	<0.001
Platelet count, $\times10^3$/uL, Median (IQR)	211.00 (162.00–277.00)	204.00 (163.00–253.00)	225.00 (172.00–303.00)	194.00 (145.00–270.00)	217.00 (156.00–296.00)	<0.001	<0.001
Prothrombin time, s, Median (IQR)	13.30 (12.20–14.60)	12.70 (11.90–13.60)	13.50 (12.50–14.70)	13.20 (12.00–14.60)	14.80 (13.15–20.55)	<0.001	<0.001
Red blood cell distribution width, %, Median (IQR)	13.80 (12.90–15.00)	13.40 (12.80–14.40)	13.40 (12.70–14.20)	15.50 (14.00–17.50)	15.10 (13.80–16.70)	<0.001	<0.001
Glucose, mg/dL, Median (IQR)	121.00 (101.00–165.00)	108.00 (95.00–127.00)	133.00 (110.00–201.00)	117.00 (98.00–153.00)	164.00 (119.00–271.75)	<0.001	<0.001
Other markers							
Oxygen saturation, %, Median (IQR)	69.00 (50.00–85.00)	65.00 (47.00–85.00)	69.00 (51.50–85.00)	69.00 (48.00–80.00)	76.50 (57.75–91.20)	0.05	0.06
BMI, kg/m^2, Median (IQR)	28.00 (25.00–33.00)	29.00 (25.00–34.00)	28.95 (25.00–33.00)	27.00 (23.00–32.00)	26.00 (23.00–31.00)	<0.001	0.73
Comorbidity, (missing = 590), *N* (%)							
Hypertension	4744 (62.35)	1238 (49.68)	1696 (60.16)	1095 (78.44)	715 (79.27)	<0.001	–
Diabetes	3104 (40.79)	666 (26.73)	1198 (42.50)	730 (52.29)	510 (56.54)	<0.001	–
Coronary artery disease	1753 (23.04)	360 (14.45)	530 (18.80)	523 (37.46)	340 (37.69)	<0.001	–
Heart failure	1132 (14.88)	176 (7.06)	286 (10.15)	430 (30.80)	240 (26.61)	<0.001	–
COPD	972 (12.77)	264 (10.59)	259 (9.19)	290 (20.77)	159 (17.63)	<0.001	–
Asthma	1091 (14.34)	392 (15.73)	372 (13.20)	232 (16.62)	95 (10.53)	<0.001	–
Cancer	1438 (18.90)	363 (14.57)	444 (15.75)	423 (30.30)	208 (23.06)	<0.001	–
Hyperlipidemia	3262 (42.87)	825 (33.11)	1169 (41.47)	779 (55.80)	489 (54.21)	<0.001	–
Obesity	3039 (37.07)	1105 (40.82)	1179 (38.69)	495 (33.31)	260 (27.11)	<0.001	–
Outcomes (60 days), *N* (%)							
Mortality	1529 (18.65)	188 (6.94)	528 (17.33)	337 (22.68)	476 (49.64)	<0.001	–
Mechanical ventilation (intubation	1154 (14.07)	190 (7.02)	527 (17.30)	195 (13.12)	242 (25.23)	<0.001	–
ICU admission	1494 (18.22)	242 (8.94)	675 (22.15)	242 (16.29)	335 (34.93)	<0.001	–

Table 2 continued

Variable	Total	Subphenotype I	Subphenotype II	Subphenotype III	Subphenotype IV	*P*-value[a]	*P*-value (age and sex adjusted)[b]
Medications, *N* (%)							
Antibiotics	2952 (36.00)	731 (27.00)	1219 (40.01)	559 (37.62)	443 (46.19)	<0.001	–
Corticosteroids	1666 (20.32)	331 (12.23)	725 (23.79)	319 (21.47)	291 (30.34)	<0.001	–
Enoxaparin	3312 (40.40)	1016 (37.53)	1582 (51.92)	418 (28.13)	296 (30.87)	<0.001	–
Heparin	1310 (15.98)	255 (9.42)	585 (19.20)	304 (20.46)	166 (17.31)	<0.001	–
Vasopressor	608 (7.42)	120 (4.43)	308 (10.11)	96 (6.46)	84 (8.76)	<0.001	–

Categories of variables were bold.

BUN blood urea nitrogen, *COPD* chronic obstructive pulmonary disease, *ESR* erythrocyte sedimentation rate, *ICU* intensive care unit, *IL-6* interleukin 6, *IQR* interquartile range, *LDH* lactate dehydrogenase.

[a] Comparisons across all 4 subphenotypes were performed using the Kruskal–Wallis test (with Dunn's test for post-hoc pairwise comparisons) or χ^2 test.

[b] *P*-values, adjusting for age and sex, were calculated by analysis of covariance (ANCOVA) was performed based on General Linear Model.

changes in clinical characteristics of the subphenotypes in the sensitivity analyses (see Supplementary Tables 6 and 7).

Subphenotypes were also rederived in the internal validation cohort, where the 4-cluster structure was found as the optimal fit as well (see Supplementary Fig. 7 and Supplementary Results). Clinical characteristics of the rederived subphenotypes in the internal validation cohort, including demographics, laboratory variables, comorbidities, and clinical outcomes, also showed very similar patterns with the subphenotypes derived in the primary analysis (see Fig. 3, Supplementary Fig. 8, and Supplementary Table 8).

To further evaluate subphenotype robustness and usability, we trained a predictive model of subphenotypes in the development cohort and used it to predict subphenotype membership in the external validation cohort. Clinical variables of presenting laboratory tests for clustering analysis were used as candidate predictors. The trained predictive model (XGBoost classifier) achieved very high performance in predicting each subphenotype (see Supplementary Fig. 9). SHapley Additive exPlanation (SHAP) values illustrated contributions of the clinical variables in distinguishing each subphenotype from others (see Supplementary Fig. 10). Patterns of the produced SHAP values were highly in line with the subphenotype characteristics: (1) normal values of the clinical variables indicated Subphenotype I; (2) abnormal inflammatory and hepatic markers were predictive of Subphenotype II; (3) abnormal renal and hematologic markers indicated were likely to indicate Subphenotype III; (4) Subphenotype IV was associated with abnormal values of most variables. After that, the trained predictive model was used to predict subphenotype memberships of patients in the external validation cohort. The predicted subphenotypes in the external validation cohort were well separated in the UMAP space (see Supplementary Fig. 11) and showed clinical characteristics similar to findings in the primary analysis (see Fig. 3, Supplementary Fig. 12, and Supplementary Table 9).

Last, results from leave-one-center-out analysis also confirmed the 4-cluster structure underlying our data (see Supplementary Fig. 13). Meanwhile, subphenotypes identified by the leave-one-center-out analysis among the whole population showed characteristics in line with those identified in the primary analysis (see Supplementary Table 10). Those above demonstrated stability of the identified subphenotypes across the five centers.

Temporal characteristics of subphenotypes

Temporal patterns of the COVID-19 subphenotypes were illustrated by the bar charts, showing the composition of subphenotype memberships of patients confirmed per week, since the outbreak in NYC, i.e., March 1, 2020 (see Fig. 4a–c). Except week 1 and week 14 that had few patients confirmed, the composition of the 4 subphenotypes per week evolved over time and showed similar patterns across the development, internal validation, and external validation cohorts. In general, patients with confirmed SARS-CoV-2 infection rapidly increased within the first month since the outbreak and reached the peak at week 5 (early April). Subphenotype I (mild symptom) and Subphenotype II (moderate symptom, low comorbidity burden) dominated the time period prior to the peak (first 4 weeks since outbreak). In contrast, Subphenotype IV (severe symptom, high comorbidity burden) had a low proportion within the first 4 weeks but showed a largely increased proportion from week 6–9. Since week 10, the proportion of Subphenotype I gradually increased while others especially Subphenotype IV shrank. Subphenotype III (moderate symptom, high comorbidity burden) had a relatively stable proportion over time.

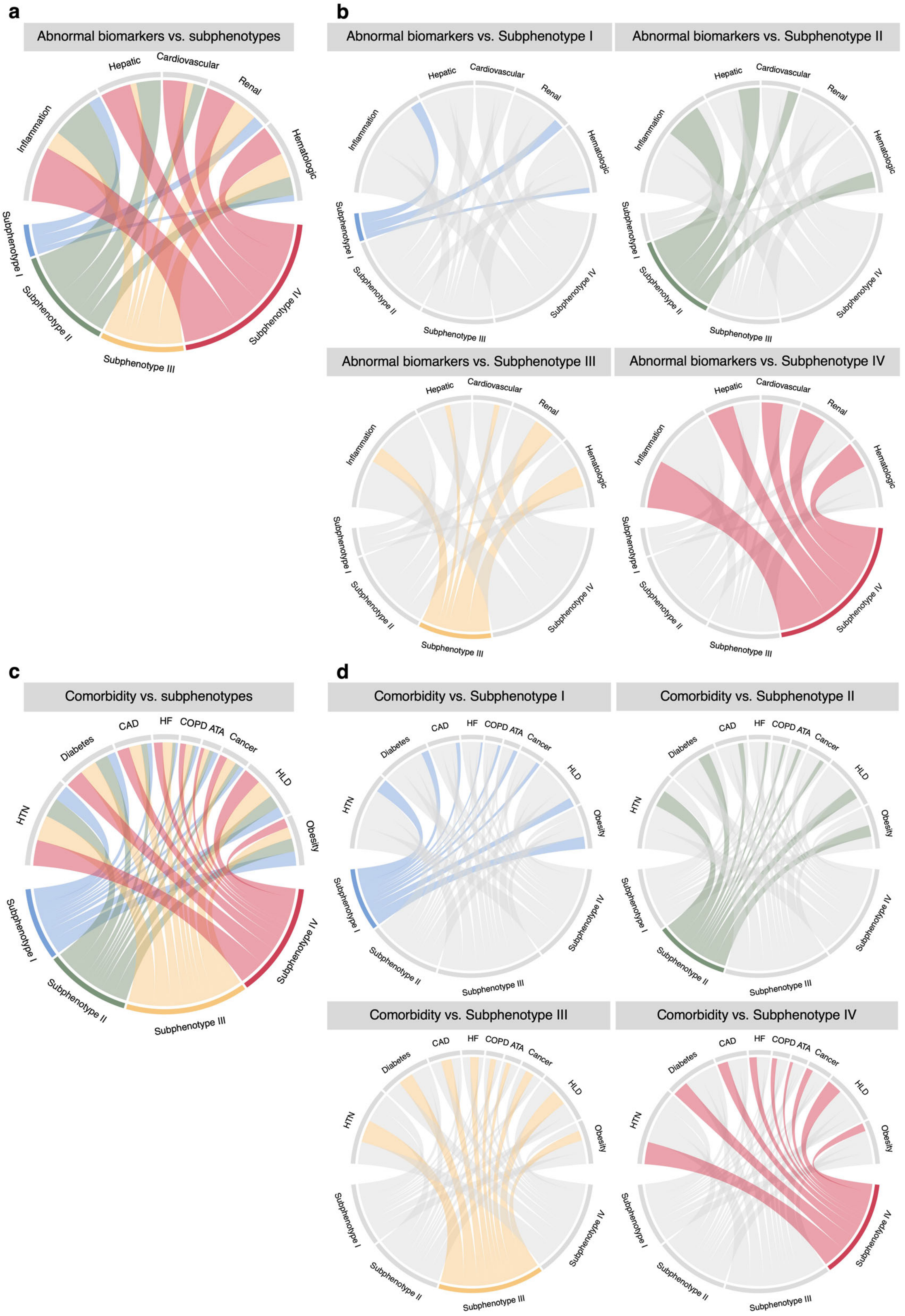

Fig. 2 Chord diagrams showing differences in abnormal clinical variables and comorbidity burden among subphenotypes. a Abnormal biomarkers vs. all subphenotypes. **b** Abnormal biomarkers vs. each subphenotype. **c** Comorbidities vs. all subphenotypes. **d** Comorbidities vs. each subphenotype. ATA asthma, CAD coronary artery disease, COPD chronic obstructive pulmonary disease, HF heart failure, HLD hyperlipidemia, HTN hypertension.

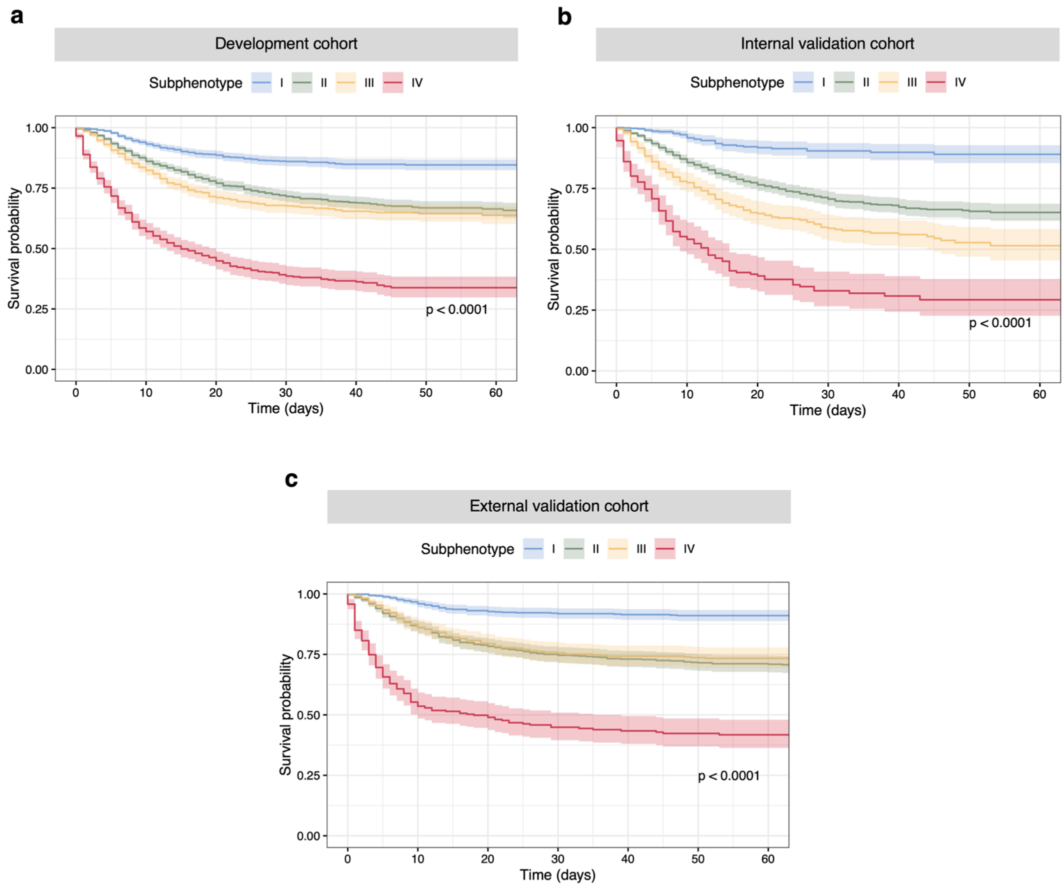

Fig. 3 Kaplan–Meier (KM) plots for 60-day mortality by subphenotypes. The survival probabilities were shown with 95% confidence interval. *X*-axis denotes time (days) after COVID-19 confirmation and *Y*-axis denotes the survival probability. **a–c** KM plots by subphenotypes in the development, internal validation, and external validation cohorts, respectively.

Impact of SDoH on subphenotypes

In general, worse SDoH in terms the socioeconomic variables were likely in Subphenotype IV (see Supplementary Table 11). Moreover, logistic regression analysis identified similar patterns of relationships between the SDoH variables with 60-day mortality risk across subphenotypes; however, absolute log odds and Hazard ratio of the SDoH variables varied across subphenotypes (see Fig. 4d and Supplementary Tables 12 and 13). For example, low absolute log odds were observed in all six SDoH variables in Subphenotype I. In contrast, we did see increased absolute log odds of all six SDoH variables in Subphenotype IV. Hazard ratio showed similar pattern.

Agglomerative hierarchical clustering based on the SDoH variables grouped the patients into a 3-cluster model (see Supplementary Fig. 14 and Supplementary Results), which can be interpreted as high (H), middle (M), and low (L) SDoH strata (see Supplementary Table 14). Stratum L, representing disadvantaged SDoH status, accounted for a slightly higher mortality rate (H vs. M vs. L, 17.59 vs. 19.91 vs.19.98%, *P*-value = 0.08). In addition, stratum L had a lower ICU admission rate (16.16%, *P*-value < 0.001). The relative high mortality but low ICU admission rate may be caused by critical care strain during periods of increased COVID-19 ICU demand, as suggested by a recent study[31]. Distributions of the SDoH strata by biological subphenotypes were shown in the Supplementary Table 15. In the analysis to further explore how SDoH strata affected the outcome of each biological subphenotype, we found varied patterns of correlations between SDoH strata and 60-day mortality (see Fig. 4e) by subphenotypes. Notably, in line with the results of the univariate analysis above, SDoH strata were likely to have a strong impact on the 60-day mortality in Subphenotype IV. Particularly, in Subphenotype IV, SDoH stratum L was associated with a 55.19% 60-mortality rate, which was 5.55% higher than the subphenotype level (49.64%, see Table 2) and 8.52% higher than that of the SDoH stratum H. In subphenotypes I, II, and III, we didn't find mortality rate discrepancy higher than 3% between any pair of SDoH strata. Similarly, considering stratum H as reference, stratum L had largely increased log odds of mortality in Subphenotype IV (log odds = 0.40, SD = 0.19, *P*-value = 0.04). (see Supplementary Table 16).

DISCUSSIONS

We derived subphenotypes of COVID-19 patients treated at five major medical centers in NYC across the whole course of the first wave of the pandemic, using the clinical data at the presentation to the emergency department (ED) or hospital. Different from the previous subphenotype studies of COVID-19[22–24], we focused on a larger, more representative, and diverse population presented at the ED and/or hospitalized without COVID-19 specific therapy. We derived subphenotypes using clustering analysis in the development cohort and validated them using a combination of multiple validation strategies, including the use of different data processing, different data filtering, and different machine learning models (both unsupervised clustering and supervised predictive models). All validation approaches confirmed the reproducibility of the 4-cluster structure of the data and clinical characteristics of the identified subphenotypes. We would also highlight that all machine learning models used for subphenotype derivation and validation were performed only on the presenting clinical variables that were routinely collected in daily patient care and

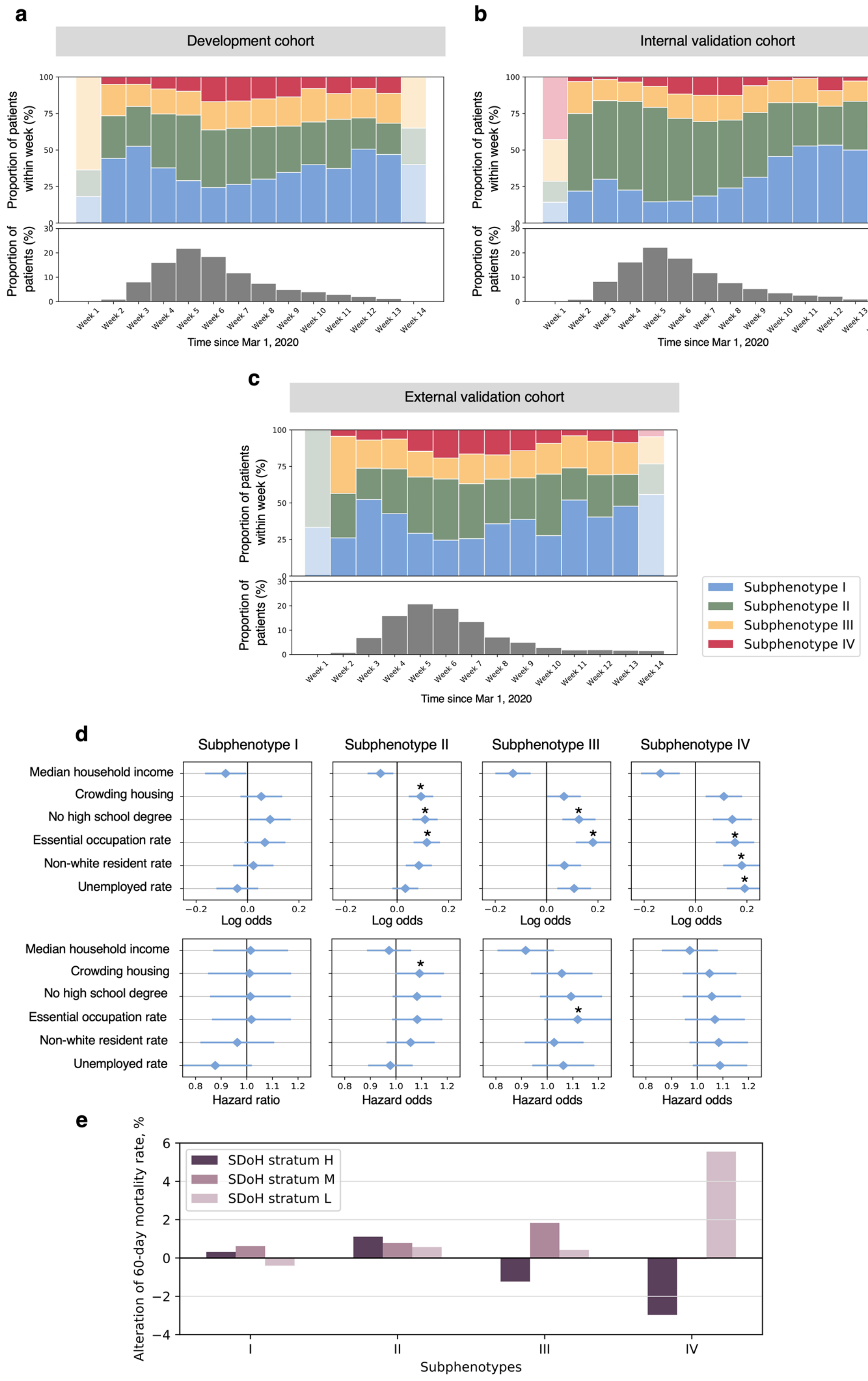

Fig. 4 Plots showing temporal patterns and SDoH implications of subphenotypes. a–c Proportions of subphenotype memberships of patients confirmed per week, since March 1, 2020. **d** Log odds and Hazard ratio (mean values and standard deviation [error bar]) showing associations between individual SDoH characteristics and 60-day mortality risk, using logistic regression analysis and Cox regression analysis, adjusting for age and sex, respectively. **e** Plot showing alteration of 60-day mortality rate (*Y*-axis) of each SDoH stratum to that of subphenotype level. **P*-value < 0.05.

are available to providers by ED or hospital admission. This allows us to potentially capture the underlying variable mechanisms of the complex disease, but also enhances the generalizability and feasibility of the identified subphenotypes to be used in clinical practices and patient enrollment in clinical trials.

Importantly, the 4 subphenotypes identified were significantly separated in demographics, clinical variables, and chronic comorbidities, and strongly predictive of the 60-day mortality outcome. Subphenotype IV included more older, male patients, abnormal markers indicating hyperinflammation, liver injury, cardiovascular problems, renal dysfunctions, and coagulation disorders, and a higher comorbidity burden (except for obesity) compared to the other subphenotypes. In contrast, Subphenotype I was composed of relatively healthy, younger females who had more normal values across all markers and comorbidity burdens compared to the other subphenotypes. There was a strong concordance between their clinical profiles and outcomes, such as Subphenotype IV showed the worst clinical outcome while Subphenotype I showed the best outcome among the 4 subphenotypes. These are in line with observations reported in a previous small cohort study[23]. Interestingly, Subphenotypes II and III showed similar, moderate-level 60-day mortality rates, but their clinical characteristic profiles suggested that they were likely to have distinct biological mechanisms. In particular, results from our primary analysis and validation approaches demonstrated that Subphenotype II was correlated with relative hyperinflammation, while Subphenotype III was associated with renal injury, lower platelet level and a high comorbidity burden (significantly higher than Subphenotypes I and II, and equivalent to Subphenotype IV). Moreover, in accordance with the clinical characteristics and outcomes, the worse subphenotypes (Subphenotypes III and IV) were more likely to receive medications in antibiotics, corticosteroids, and vasopressor than the others. These findings suggested that our identified subphenotypes offer insight into the varied mechanisms of COVID-19.

Typically, data-driven approaches for the identification of subphenotypes of human disease are based on the unsupervised clustering methods[12,14–16,22–24,32]. The natural attributes of the unsupervised methodology in discovering underlying patterns from data make them the best fit for subphenotype identification. Once the subphenotypes were determined, there would be a need of subphenotype membership assignments for new patients. However, previous studies barely discussed such down-stream usability of the identified subphenotypes. In this analysis, we built a supervised predictive model of the identified subphenotypes. Our predictive model achieved an ideal prediction performance in the development cohort and predicted subphenotypes in the external validation cohort that presented the same pattern of clinical characteristics with that of the originally derived subphenotypes. In this way, instead of validating the subphenotypes in a different route, the predictive model brought additional implications as: (1) it provides a feasible and accurate way to apply the identified subphenotypes to clinical practice; and (2) contributions of the clinical variables in subphenotype prediction calculated by the SHAP method showed concordant patterns with the subphenotypes' clinical characteristics and hence confirmed biological profiles of the subphenotypes in the multivariate prospective.

Time is a crucial factor in the spread of COVID-19. Previous studies have examined the temporal trends of COVID-19 outcomes such as in-hospital mortality rate during the course of the pandemic[26,27], but limited attention has been drawn on evolving patterns of COVID-19 phenotypes. We filled this gap in the present study. Our observations suggested varied temporal trends of the identified subphenotypes during the first 14 weeks of the pandemic in NYC. Interestingly, since the COVID-19 outbreak in NYC on March 1, 2020, Subphenotypes I and II dominated the time period prior to the peak (first 4 weeks since outbreak), possibly as they contained more relatively younger patients who may have had more frequent social activities to be infected. Subphenotype IV, with older age, worse health conditions, and poorer outcomes, was boosted within the second month (April 2020) post spread peak, consistent with tremendous mortality rate of NYC in April[33].

This would suggest that younger, biologically strong patients (Subphenotypes I and II) got infections early and boosted the spread, while older, biologically vulnerable patients (Subphenotype IV) accounted for the second infections within a population probably due to housing. After that, the proportion of Subphenotype I out of all patients confirmed per week gradually expanded while that of the others, especially Subphenotype IV shrank. The potential reason would be that valuable experience (such as the improved use of masks and social distancing), reinforced healthcare systems, and announced health policies did protect the population who likely develop severe subphenotypes (Subphenotype IV). In general, such temporal trends of the biological subphenotypes would be a considerable, fine-grained explanation of the observed outcome (mortality rate) evolving trends in epidemiology[26].

SDoH such as vulnerable socioeconomic neighborhood status have been associated with poor outcomes of COVID-19[26,30]. In this work, we explored the impact of SDoH on different biological subphenotypes from both univariate and multivariate perspectives. We first examined the associations of individual socioeconomic characteristics with mortality risk in each subphenotype. We then derived comprehensive SDoH strata using the data-driven clustering method and evaluated their correlations with mortality risk in each subphenotype. The results confirmed our hypothesis that SDoH impacts biological subphenotypes differently. The highly expanded mortality risk log odds of individual SDoH variables and discrepancy of mortality rate among SDoH strata indicate that SDoH has a much stronger association with mortality outcomes in Subphenotype IV, compared to the others. In other words, once a sick, elderly patient shows up with COVID-19 (Subphenotype IV), the disadvantaged socioeconomic status significantly increased their mortality. In contrast, disadvantaged SDoH status was unlikely to lead to significantly increased mortality risk in Subphenotype I. This evidence further demonstrated that the COVID-19 pandemic has disproportionately affected patients with lower socioeconomic status. In general, our findings added new information on social disparities in the COVID-19 pandemic. Unlike previous studies[29,30,34,35] that focused on the entire population, we extended the study from a new angle by focusing on the biologically different populations (i.e., subphenotypes). Our findings also showed evidence that the identified subphenotypes would provide considerable guidance in health policy to reduce social disparities in the pandemic.

While this study presents a new contribution in the efforts to parse the biological heterogeneity of COVID-19, there remain several limitations. First of all, our data-driven approach relied on the availability of patient data. In this study, we identified subphenotypes using the routinely collected clinical variables that were correlated with COVID-19[36] and available in the INSIGHT database[37]. We were not able to extract presenting symptoms and vital data while the incorporation of such data would add in new insights. Second, in our study, the analyzed data were collected at ED or hospital presentation, so the time between COVID-19 symptom onset to ED or hospital presentation could be a covariate of disease severity and clinical outcomes. However, such data was not available in the INSIGHT database. Third, missing values may affect the robustness of the identified subphenotypes. In order to address this issue, we excluded variables with high missingness. For the remaining variables, we used the K-nearest neighbors imputation algorithm[38]. Even so, we still missed these real values hence may incorporate bias. Fourth, our study was based on presenting clinical data, such that each patient was characterized in a snapshot. The full use of longitudinal data of patients may allow us to capture the complexity of the disease arc to identify interesting subphenotypes. Previous studies tried to derive COVID-19 subphenotypes based on longitudinal information[22,24], yet they were based on univariate trajectory data in small cohorts. The collection of multivariate, longitudinal data in large cohorts remains challenging and modeling such data to identify subphenotypes requires improved data-driven methods[12,13,21]. Fifth, this is a multiple institutional analysis in NYC. To evaluate the generalizability of the identified subphenotypes, further validation on data collected from other areas is needed in future work.

METHODS

Study design and cohort description

We used data of COVID-19 patients from INSIGHT Clinical Research Network (CRN)[37]. INSIGHT is funded by the Patient-Centered Outcomes Research Institute (PCORI) and aggregates clinical data of diverse patient populations across five academic medical centers in New York City (NYC), including New York University Langone Medical Center (NYU-LMC), New York Presbyterian—Weill Cornell Medical Center (NYP-WCMC), New York Presbyterian—Columbia University Medical Center (NYP-CUMC), Mount Sinai Health System (MSHS), and Montefiore Medical Center (MMC). COVID-19 diagnosis was defined as having at least one positive laboratory test result for SARS-CoV-2 infection or at least one ICD-10 diagnosis code for COVID-19 (see Supplementary Note 1). Study participants were adult patients who were diagnosed with COVID-19 and treated in ED or inpatient settings in these five health centers from March 1 to June 12, 2020. Criteria used to assess patient eligibility are illustrated in the Supplementary Fig. 1. Exclusion criteria include younger than 18 years old; duplicated patient IDs; having no emergency department (ED) or inpatient (IP) admission within 14 days after COVID-19 confirmation; or having missing values on all clinical variables. Considering the population diversity of the five medical centers (see Supplementary Table 1), we combined patients of four centers and randomly divided them into the development cohort (70%) and internal validation cohort (30%). Patients of the last center were used as the external validation cohort.

Candidate variables for subphenotype identification

We considered 30 clinical variables associated with COVID-19 onset, symptoms, or outcomes[36] and available in the INSIGHT database as the candidate variables to derive subphenotypes. The variables included inflammatory markers (C-reactive protein, erythrocyte sedimentation rate [ESR], interleukin 6 [IL-6], procalcitonin, bands [i.e., premature neutrophil], lactate dehydrogenase [LDH], lymphocyte count, neutrophil count, and white blood cell count), inflammatory and hepatic markers (albumin and ferritin), hepatic markers (alanine aminotransferase [ALT], aspartate aminotransferase [AST], and bilirubin), markers of cardiovascular conditions (creatine kinase [CK], lactate, troponin I, and troponin T), markers of renal dysfunctions (bicarbonate, blood urea nitrogen [BUN], creatinine, chloride, and sodium), markers of hematologic dysfunctions (d-dimer, hemoglobin, platelet count, prothrombin time [PT], red blood cell distribution width [RDW], and glucose), and oxygen saturation. For each patient, we extracted the first value of each clinical variable within the collection window, which was defined as: (1) time period from COVID-19 confirmation to the first inpatient encounter, if the patient has an inpatient admission within 14 days after confirmation; or (2) 14 days after COVID-19 confirmation if there was only ED encounters but no inpatient admissions following the COVID-19 diagnosis. If there was no record in the collection window, we extracted the last value within 3 days before confirmation (see Supplementary Fig. 2).

Other clinical characteristics, clinical outcomes, and medications

We also examined other clinical characteristics of the patients, including demographics, comorbidities, and body mass index (BMI). Demographics included age, sex, and race. Baseline comorbidities included hypertension, diabetes, coronary artery disease (CAD), heart failure, chronic obstructive pulmonary disease (COPD), asthma, cancer, obesity, and hyperlipidemia. For each patient, the most recent BMI data were collected. We analyzed 60-day all-cause mortality as the primary outcome for the patients. Need for mechanical ventilation and admission to the intensive care unit (ICU) were the secondary outcomes. We also analyzed the treatments for COVID-19, including antibiotics (combining ceftriaxone, azithromycin, piperacillin tazobactam, meropenem, vancomycin, and doxycycline), corticosteroids (combining prednisone, methylprednisolone, dexamethasone, and hydrocortisone), hydroxychloroquine, enoxaparin, heparin, and vasopressor. These above data were collected from patient records available in the INSIGHT database as well.

SDoH data

To explore the impact of SDoH status on the subphenotypes, we extracted patients' neighborhood socioeconomic characteristics, including median household income, percentage of residents without a high school degree, percentage of residents who are essential workers, percentage of households with crowding housing conditions (i.e., households with >1 person per room), percentage of non-white residents, and unemployment rate. These characteristics were extracted from the 2018 American Community Survey[39]. Previous studies[40–46] have indicated that these social conditions are associated with higher probability of infection, hospitalization, and other adverse outcomes, e.g., mortality, in COVID-19.

Data preparation

We first assessed the value distributions and missingness of the 30 candidate clinical variables (see Supplementary Tables 2 and 3). For data quality control, seven variables of high missingness (missing more than 70% values) were excluded and the remaining 23 variables were used for deriving subphenotypes. The seven high-missing variables were included in sensitivity analysis. Details of usage of the clinical variables were presented in Supplementary Table 4. Logarithmic transformation was applied to the non-normal distributed variables (see Supplementary Table 4). In order to eliminate the effects of value magnitude, all variables were scaled based on z-score.

Last, K-nearest neighbors (KNN) imputation[38] was used to address missing values. More details can be found in the Supplementary Note 2.

Subphenotype derivation, validation, and prediction

We originally derived subphenotypes using the development cohort. More specifically, agglomerative hierarchical clustering with Euclidean distance calculation and Ward linkage criterion[47] was applied to the 23 clinical variables after data preparation. We used agglomerative hierarchical clustering because it is robust to different types of data distributions and typically produces a dendrogram that visualizes data structure to help determine the optimal cluster number. Besides dendrogram, we calculated 21 measures of clustering models provided by 'NbClust' software[48] to determine the optimal number of clusters, i.e., subphenotypes. More details can be found in the Supplementary Note 3.

In order to evaluate the reproducibility, we validated our subphenotypes in four ways. First, we performed sensitivity analyses using the development cohort to evaluate (1) sensitivity to quality control and outliers and (2) sensitivity to clustering algorithms. To assess sensitivity to quality control and outliers, we incorporated all 30 candidate variables and excluded patients who have outlier values, defined as values out of the range of $[\mu - 5\sigma, \mu + 5\sigma]$, where μ and σ are the mean and standard deviation of the specific variable. Then similar to the primary analysis, we performed agglomerative hierarchical clustering to rederive subphenotypes and determined optimal cluster number using dendrogram and 'NbClust'. To assess sensitivity to clustering algorithms, we rederived subphenotypes using the Gaussian mixture model (GMM)[49], which is a probabilistic model for clustering analysis based on a mixture of Gaussian distributions. The optimal cluster number in GMM was determined by comprehensively considering Akaike information criterion (AIC), Bayesian information criterion (BIC), and median probability of group membership (see Supplementary Note 4).

Second, we used the internal validation cohort and rederived subphenotypes using the same agglomerative hierarchical clustering with the primary analysis for validation. The optimal cluster number was determined using dendrogram and 'NbClust' as well.

Third, for the aims of confirming subphenotypes and their usability, we used the supervised predictive model. More specifically, considering subphenotype membership of each patient as the label to predict, we built a predictive model of subphenotypes based on the 23 clinical variables used for subphenotype derivation. The predictive model was based on the supervised XGBoost classifier[50], a powerful tree-based machine learning model. The predictive model was trained in the development cohort using a 10-fold cross-validation strategy. To address the multi-label classification (since we identified more than 2 subphenotypes), a one-vs-the-rest strategy was used in model training. Prediction performance was measured by receiver operating characteristics curve (ROC) and area under ROC curve (AUC). We also engaged the SHapley Additive exPlanation (SHAP) values to assess contributions of the clinical variables in distinguishing each subphenotype from the others. Once the predictive model was trained, it was performed on the external validation cohort to predict the patients' subphenotype memberships.

To assess stability of the subphenotypes across the five medical centers, we further performed leave-one-center-out analysis. Specifically, within each loop of the leave-one-center-out procedure, we used a specific center as the within-loop validation cohort. The remaining four centers were combined as the within-loop development cohort, which was used to derive clusters, i.e., subphenotypes, and train XGboost-based predictive model of subphenotypes.

Then the predictive model was used to predict patients' cluster membership in the within-loop validation cohort. Similar to the consensus clustering[51], after all the five iterations, we generated a $N \times N$ cluster consensus matrix $M = [m_{ij}]$, where N is the total number of patients and m_{ij} is the 'consensus value' between patients i and j, which is defined as the frequency that the two patients are assigned to the same cluster during the leave-one-center-out procedure. Typically, the consensus value ranges from 0 to 1, where a value of 0 means that the pair of patients never been grouped to the same cluster, while a value of 1 means that they always assigned to the same cluster during the leave-one-center-out procedure. Finally, the agglomerative hierarchical clustering was performed on the consensus matrix. Our goal is to confirm if we can still obtain cluster structure observed in our primary analysis under the leave-one-center-out procedure.

Subphenotype interpretation

For the aim of subphenotype interpretation, we first visualized the subphenotypes in two ways: (1) 2-D visualization calculated by Uniform Manifold Approximation and Projection (UMAP) algorithm[52] based on clinical variables for clustering (showing distributions of subphenotypes within low-dimensional space); (2) chord diagrams[53] showing differences of subphenotypes in terms of abnormal clinical variable groups and comorbidities. More details can be found in the Supplementary Note 5.

We also characterized subphenotypes by evaluating their differences in demographics, all clinical variables, comorbidities, clinical outcomes, and medications prescribed after COVID-19 confirmation. Data were presented as median (interquartile range [IQR]) for continuous variables and exact patient number (percentage) for categorical variables. To compare subphenotypes, we performed the Kruskal–Wallis test for continuous data and χ^2 test for categorical data. Analysis of covariance (ANCOVA) was also applied for between-subphenotypes comparisons, adjusting for age and gender. Two-tailed P-values smaller than 0.05 were considered as the threshold for statistical significance. Survival analyses were performed to assess associations of subphenotypes to clinical outcomes, where Kaplan–Meier plots were created accordingly.

Temporal pattern of subphenotypes

To evaluate the temporal pattern of the subphenotypes during the course of the pandemic, we created bar charts to visualize the proportion of each subphenotype out of the total patients confirmed per week, since the COVID-19 outbreak in NYC (March 1, 2020).

Impacts of SDoH on COVID-19 subphenotypes

Multiple analyses were conducted to assess the impact of SDoH on COVID-19 subphenotypes. For each subphenotype, we first performed logistic regression analysis and Cox regression analysis to assess the association of each SDoH variable with 60-day mortality, adjusting for age, sex, and/or clinical variables. After that, we performed agglomerative hierarchical clustering on the six socioeconomic variables to derive comprehensive SDoH strata. Within each subphenotype, we compared 60-day mortality rates between the SDoH strata. We also used logistic regression analysis and Cox regression analysis to assess the association of SDoH strata with 60-day mortality, adjusting for age and sex, within each subphenotype.

Ethics

The Institutional Review Board of the Weill Cornell Medicine approved this study (Protocol number: 20-04021948). In this study, because all clinical data obtained from the INSIGHT clinical research network were deidentified, a waiver of consent was obtained from the ethics institution.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

All clinical data studied in this work can be downloaded from INSIGHT clinical research network at https://insightcrn.org/our-data/, via request. Neighborhood socioeconomic data were obtained from the 2018 American Community Survey that can be downloaded from the United States Census Bureau at https://data.census.gov/cedsci/.

CODE AVAILABILITY

All computer codes in this study are available at https://github.com/ChangSu10/COVID-Insight-subphenotyping. Implementation of our work is based on Python 3.7 and R 3.6. More specifically, clustering models were implemented based on Python packages 'scikit-learn 0.23.2' (https://scikit-learn.org/stable/) and 'scipy 1.5.3' (https://www.scipy.org). Supervised predictive modeling was based on 'XGBoost 1.2.1' (https://xgboost.readthedocs.io/en/latest/) and 'SHAP 0.35.0' (https://shap.readthedocs.io/en/latest/). Data dimension reduction and visualization were performed based on Python package 'UMAP-learn 0.3.9' (https://umap-learn.readthedocs.io/en/latest/). R package 'NbClust' (https://cran.r-project.org/web/packages/NbClust/NbClust.pdf) was used to calculate measures of clusters to determine the optimal cluster number in agglomerative hierarchical clustering. Chord diagrams were created using R package 'circlize' (https://cran.r-project.org/web/packages/circlize/index.html). All statistical tests and survival analyses were performed based on R.

Received: 9 March 2021; Accepted: 21 June 2021;
Published online: 14 July 2021

REFERENCES

1. Zhu, N. et al. A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).
2. Richardson, S. et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City Area. *JAMA* **323**, 2052–2059 (2020).
3. Tabata, S. et al. Clinical characteristics of COVID-19 in 104 people with SARS-CoV-2 infection on the Diamond Princess cruise ship: a retrospective analysis. *Lancet Infect. Dis.* **20**, 1043–1050 (2020).
4. Desai, N. et al. Temporal and spatial heterogeneity of host response to SARS-CoV-2 pulmonary infection. *Nat. Commun.* **11**, 6319 (2020).
5. Wiersinga, W. J., Rhodes, A., Cheng, A. C., Peacock, S. J. & Prescott, H. C. Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): a review. *JAMA* **324**, 782–793 (2020).
6. Takahashi, T. et al. Sex differences in immune responses that underlie COVID-19 disease outcomes. *Nature* **588**, 315–320 (2020).
7. Gupta, S. et al. Association between early treatment with tocilizumab and mortality among critically Ill patients with COVID-19. *JAMA Intern. Med.* **181**, 41–51 (2021).
8. Domecq, J. P. et al. Outcomes of patients with coronavirus disease 2019 receiving organ support therapies: The International Viral Infection and Respiratory Illness Universal Study Registry. *Crit. Care Med.* **49**, 437–448 (2021).
9. The RECOVERY Collaborative Group. Dexamethasone in hospitalized patients with Covid-19–preliminary report. *N. Engl. J. Med* **384**, 693–704 (2020).
10. Kreuziger, L. B. et al. *COVID-19 and VTE/Anticoagulation: Frequently Asked Questions* (The American Society of Hematology, 2021).
11. National Heart, Lung, and Blood Institute (NHLBI). *Full-dose blood thinners decreased need for life support and improved outcome in hospitalized COVID-19 patients* (National Institutes of Health (NIH), 2021).
12. Weng, C., Shah, N. H. & Hripcsak, G. Deep phenotyping: embracing complexity and temporality-towards scalability, portability, and interoperability. *J. Biomed. Inform.* **105**, 103433 (2020).
13. Reddy, K. et al. Subphenotypes in critical care: translation into clinical practice. *Lancet Respir. Med.* **8**, 631–643 (2020).
14. Seymour, C. W. et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA* **321**, 2003–2017 (2019).
15. Bhavani, S. V. et al. Identifying novel sepsis subphenotypes using temperature trajectories. *Am. J. Respir. Crit. Care Med.* **200**, 327–335 (2019).
16. Calfee, C. S. et al. Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. *Lancet Respir. Med.* **2**, 611–620 (2014).
17. Ahmad, T. et al. Clinical implications of chronic heart failure phenotypes defined by cluster analysis. *J. Am. Coll. Cardiol.* **64**, 1765–1774 (2014).
18. Cikes, M. et al. Machine learning-based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy. *Eur. J. Heart Fail.* **21**, 74–85 (2019).
19. Li, L. et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci. Transl. Med.* **7**, 311ra174 (2015).
20. Neff, R. A. et al. Molecular subtyping of Alzheimer's disease using RNA sequencing data reveals novel mechanisms and targets. *Sci. Adv.* **7**, eabb5398 (2021).
21. Bos, L. D. J., Sinha, P. & Dickson, R. P. The perils of premature phenotyping in COVID-19: a call for caution. *Eur. Respir. J.* **56**, 2001768 (2020).

22. Bhavani, S. V., Huang, E. S., Verhoef, P. A. & Churpek, M. M. Novel temperature trajectory subphenotypes in COVID-19. *Chest* **158**, 2436–2439 (2020).
23. Legrand, M. et al. Differences in clinical deterioration among three subphenotypes of COVID-19 patients at the time of first positive test: results from a clustering analysis. *Intensive Care Med.* **47**, 113–115 (2021).
24. Su, C. et al. Identifying organ dysfunction trajectory-based subphenotypes in critically ill patients with COVID-19. *medRxiv*. https://www.medrxiv.org/content/10.1101/2020.07.16.20155382v3 (2020).
25. Rodríguez, A. et al. Deploying unsupervised clustering analysis to derive clinical phenotypes and risk factors associated with mortality risk in 2022 critically ill patients with COVID-19 in Spain. *Crit. Care* **25**, 63 (2021).
26. Asch, D. A. et al. Variation in US hospital mortality rates for patients admitted with COVID-19 during the first 6 months of the pandemic. *JAMA Intern. Med.* **181**, 471–478 (2021).
27. Jorge, A. et al. Temporal trends in severe COVID-19 outcomes in patients with rheumatic disease: a cohort study. *Lancet Rheumatol.* **3**, e131–e137 (2021).
28. Gray, D. M., Anyane-Yeboa, A., Balzora, S., Issaka, R. B. & May, F. P. COVID-19 and the other pandemic: populations made vulnerable by systemic inequity. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 520–522 (2020).
29. Wadhera, R. K. et al. Variation in COVID-19 hospitalizations and deaths across New York City Boroughs. *JAMA* **323**, 2192–2195 (2020).
30. Azar, K. M. J. et al. Disparities in outcomes among COVID-19 patients in a large Health Care System In California. *Health Aff. (Millwood)* **39**, 1253–1262 (2020).
31. Bravata, D. M. et al. Association of Intensive Care Unit Patient Load and demand with mortality rates in US Department of Veterans Affairs Hospitals during the COVID-19 pandemic. *JAMA Netw. Open* **4**, e2034266–e2034266 (2021).
32. Knox, D. B., Lanspa, M. J., Kuttler, K. G., Brewer, S. C. & Brown, S. M. Phenotypic clusters within sepsis-associated multiple organ dysfunction syndrome. *Intensive Care Med.* **41**, 814–822 (2015).
33. Thompson, C. N. et al. COVID-19 Outbreak—New York City, February 29-June 1, 2020. *Morbidity Mortal. Wkly. Rep.* **69**, 1725–1729 (2020).
34. Wang, Z. et al. Analysis of hospitalized COVID-19 patients in the Mount Sinai Health System using electronic medical records (EMR) reveals important prognostic factors for improved clinical outcomes. *medRxiv*. https://www.medrxiv.org/content/10.1101/2020.04.28.20075788v1 (2020).
35. Federgruen, A. & Naha, S. Variation in Covid-19 cases across New York City. *medRxiv*. https://www.medrxiv.org/content/10.1101/2020.05.25.20112797v2 (2020).
36. Wynants, L. et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* **369**, m1328 (2020).
37. Kaushal, R. et al. Changing the research landscape: the New York City Clinical Data Research Network. *J. Am. Med. Inform. Assoc.* **21**, 587–590 (2014).
38. Troyanskaya, O. et al. Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001).
39. The Census Bureau. *American Community Survey: 2018 Data Release New and Notabls.* https://www.census.gov/programs-surveys/acs/news/data-releases/2018/release.html.
40. Kolak, M., Bhatt, J., Park, Y. H., Padrón, N. A. & Molefe, A. Quantification of neighborhood-level social determinants of Health in the Continental United States. *JAMA Netw. Open* **3**, e1919928–e1919928 (2020).
41. Whittle, R. S. & Diaz-Artiles, A. An ecological study of socioeconomic predictors in detection of COVID-19 cases across neighborhoods in New York City. *BMC Med.* **18**, 271 (2020).
42. von Seidlein, L., Alabaster, G., Deen, J. & Knudsen, J. Crowding has consequences: prevention and management of COVID-19 in informal urban settlements. *Build. Environ.* **188**, 107472 (2021).
43. Hawkins, R. B., Charles, E. J. & Mehaffey, J. H. Socio-economic status and COVID-19–related cases and fatalities. *Public Health* **189**, 129–134 (2020).
44. Lieberman-Cribbin, W., Tuminello, S., Flores, R. M. & Taioli, E. Disparities in COVID-19 testing and positivity in New York City. *Am. J. Prev. Med.* **59**, 326–332 (2020).
45. Do, D. P. & Frank, R. Unequal burdens: assessing the determinants of elevated COVID-19 case and death rates in New York City's racial/ethnic minority neighbourhoods. *J. Epidemiol. Community Health* **75**, 321 (2021).
46. Hong, B., Bonczak, B. J., Gupta, A., Thorpe, L. E. & Kontokosta, C. E. Exposure density and neighborhood disparities in COVID-19 infection risk. *Proc. Natl Acad. Sci. USA* **118**, e2021258118 (2021).
47. Murtagh, F. & Legendre, P. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *J. Classif.* **31**, 274–295 (2014).
48. Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. NbClust: an R package for determining the relevant number of clusters in a data set. *J. Stat. Softw.* **61**, 1–36 (2014).
49. Reynolds, D. A. Gaussian mixture models. *Encycl. Biometrics* **741**, 659–663 (2009).
50. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, 2016).
51. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* **52**, 91–118 (2003).
52. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *J. Open Source Softw.* **3**, 861 (2018).
53. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).

ACKNOWLEDGEMENTS

This study is funded by the COVID-19-Related Project Enhancement to the grant PCORI/HSD-1604-35187 ("Identifying and Predicting Patients with Preventable High Utilization", PI: Kaushal) from the Patient-Centered Outcomes Research Institute (PCORI).

AUTHOR CONTRIBUTIONS

E.S. and F.W. for conceptualization, investigation, writing, reviewing, and editing of the manuscript. C.S. for investigation, data analysis, drafting, editing, and reviewing manuscript. M.G.W. for providing data support, discussion, and commenting the manuscript. Y.Z., J.H.F., and R.K. for discussion, commenting, and editing the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-021-00481-w.

Correspondence and requests for materials should be addressed to R.K., E.J.S. or F.W.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.