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

Authors: Chang Su, PhD[1]; Yongkang Zhang, PhD[1]; James H Flory[2], MD; Mark G. Weiner, MD[1]; Rainu Kaushal*, MD[1,3]; Edward J. Schenck*, MD[3,4]; Fei Wang*, PhD[1]

Affiliations:
[1]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY;
[2]Memorial Sloan-Kettering Cancer Center, New York, NY;
[3]New York-Presbyterian Hospital, Weill Cornell Medicine, New York, NY;
[4]Division of Pulmonary & Critical Care Medicine, Joan and Sanford I. Weill Department of Medicine, Weill Cornell Medicine, New York, NY
*Corresponding authors

Abstract

The coronavirus disease 2019 (COVID-19) is heterogeneous and our understanding of the biological mechanisms of host response to the novel 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, four 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 re-derivation 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 new insights to health policy to reduce social disparity in the pandemic.

[Main text]

The outbreak of coronavirus disease 2019 (COVID-19), caused by the novel 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[22-24] 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[25,26] and according to social determinants of health (SDoH)[27-29]. 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 pre-processing, 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 Figure 1.

Results

Patients

A total of 14418 patients with confirmed COVID-19 between March 1st and June 12th 2020, treated in ED or inpatient settings, were included for analysis from the five major 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), Mount Sinai Health System (MSHS), Montefiore Medical Center (MMC), and New York Presbyterian - Columbia University Medical Center (NYP-CUMC). Details of inclusion and exclusion criteria are presented in eFigure 1 in Supplement. Considering population diversity (especially race) of the five medical centers (see eTable 1 in Supplement), we combined NYU-LMC, NYP-WCMC,

MSHS, and MMC and randomly divided them into the development cohort (70%) and internal validation cohort (30%); patients of NYP-CUMC were used as the external validation cohort (see Figure 1 and eFigure. 1 in Supplement).

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 eResults, eFigures 3 and 4 in Supplement). Characteristics including demographics, clinical variables, comorbidities, clinical outcomes, and medication treatments across the 4 subphenotypes were presented in Table 2 and Figures 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. 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 eResults, eFigures 3 and 4, and eTable 5 in Supplement). Patients' memberships of the 4 clusters re-derived by sensitivity analyses were highly consistent with those derived in the primary analysis (see eFigure 6 in Supplement). Moreover, we did not find substantial changes in clinical characteristics of the subphenotypes in the sensitivity analyses (see eTables 6 and 7 in Supplement).

Subphenotypes were also re-derived in the internal validation cohort, where the 4-cluster structure was found as the optimal fit as well (see eResults and eFigure7 in Supplement). Clinical characteristics of the re-derived 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 Figure 3, and eTable 8 and eFigure 8 in Supplement).

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 eFigure 9 in Supplement). SHapley Additive exPlanation (SHAP) values illustrated contributions of the clinical variables in distinguishing each subphenotype from others (see eFigure 10 in Supplement). 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 eFigure 11 in Supplement) and showed clinical characteristics similar to findings in the primary analysis (see Figure 3, and eTable 9 and eFigure 12 in Supplement).

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 Figures 4a-c). Except week 1 and week 14 that had few patients confirmed, the composition of the four 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 to week 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.

Impact of SDoH on subphenotypes

In general, worse SDoH in terms the socioeconomic variables were likely related to Subphenotype IV (see eTable 10 in Supplement). Moreover, logistic regression analysis identified similar patterns of relationships between the SDoH variables with 60-day mortality risk across subphenotypes; however, absolute log odds of the SDoH variables varied across subphenotypes (see Figure 4d and eTable 11 in Supplement). 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.

Agglomerative hierarchical clustering based on the SDoH variables grouped the patients into a 3-cluster model (see eResults and eFigure 13 in Supplement), which can be interpreted as high (H), middle (M), and low (L) SDoH strata (see eTable 12 in Supplement). 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[30]. Distributions of the SDoH strata by biological subphenotypes were shown in eTable 13 in Supplement. 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 Figure 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 eTable 13 in Supplement)

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 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 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[12,14-16,22-24,31]. 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 novel 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[25,26], 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[32].

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 health care systems, and announced health policies did protect the population who likely develop severe subphenotypes (Subphenotype IV). In general, such temporal trends of the novel biological subphenotypes would be a considerable, fine-grained explanation of the observed outcome (mortality rate) evolving trends in epidemiology[25].

SDoH such as vulnerable socioeconomic neighborhood status have been associated with poor outcomes of COVID-19[25,29]. 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[28,29,33,34] 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.

Limitations

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[35] and available in the INSIGHT database[36]. 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 state-of-the-art K-nearest neighbors imputation algorithm[37]. 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.

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. *The Lancet Infectious Diseases* **20**, 1043-1050 (2020).
4. Desai, N.*, et al.* Temporal and spatial heterogeneity of host response to SARS-CoV-2 pulmonary infection. *Nature Communications* **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 Internal Medicine* **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.* **Online First**(9000).
9. Dexamethasone in Hospitalized Patients with Covid-19 — Preliminary Report. *N. Engl. J. Med.* (2020).
10. Kreuziger, L.B.*, et al.* COVID-19 and VTE/Anticoagulation: Frequently Asked Questions. 2021 (The American Society of Hematology, accessed by Feb 8, 2021).
11. Full-dose blood thinners decreased need for life support and improved outcome in hospitalized COVID-19 patients. 2021 (National Institutes of Health (NIH), accessed by Feb 10, 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. *The Lancet Respiratory Medicine* **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. *The Lancet Respiratory Medicine* **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. *Science Advances* **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 sub-phenotypes 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*, 2020.2007.2016.20155382 (2020).
25. 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 Internal Medicine* (2020).
26. Jorge, A.*, et al.* Temporal trends in severe COVID-19 outcomes in patients with rheumatic disease: a cohort study. *The Lancet Rheumatology* **3**, e131-e137 (2021).
27. 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. *Nature Reviews Gastroenterology & Hepatology* **17**, 520-522 (2020).
28. Wadhera, R.K.*, et al.* Variation in COVID-19 Hospitalizations and Deaths Across New York City Boroughs. *JAMA* **323**, 2192-2195 (2020).
29. 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).
30. 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 Network Open* **4**, e2034266-e2034266 (2021).
31. 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).
32. Thompson, C.N.*, et al.* COVID-19 Outbreak - New York City, February 29-June 1, 2020. *Morbidity and mortality weekly report (MMWR)* **69**, 1725-1729 (2020).
33. 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*, 2020.2004.2028.20075788 (2020).
34. Federgruen, A. & Naha, S. Variation in Covid-19 Cases Across New York City. *medRxiv*, 2020.2005.2025.20112797 (2020).
35. Wynants, L.*, et al.* Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* **369**, m1328 (2020).
36. 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).
37. Troyanskaya, O.*, et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520-525 (2001).

Online Methods

Study design and cohort description

We used data of COVID-19 patients from INSIGHT Clinical Research Network (CRN)[36]. 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 eMethods in Supplement). 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 eFigure 1 in Supplement. 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 eTable 1 in Supplement), we combined patients of NYU, NYP-WCMC, MSHS, and MMC and randomly divided them into the development cohort (70%) and internal validation cohort (30%). Patients of NYP-CUMC 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[37] 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 eFigure 2 in Supplement).

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 was 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.

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[38]. Previous studies have indicated that these social conditions are associated with higher probability of infection, hospitalization, and other adverse outcomes.

Statistical methods

Data preparation

We first assessed the value distributions and missingness of the 30 candidate clinical variables (see eTables 2 and 3 in Supplement). For data quality control, 7 variables of high missingness (missing more than 70% values) were excluded and the remaining 23 variables were used for deriving subphenotypes. Logarithmic transformation was applied to the non-normal distributed variables (see eTable 4 in Supplement). In order to eliminate the effects of value magnitude, all

variables were scaled based on z-score. K-nearest neighbors (KNN) imputation[39] was used to address missing values (see eMethods in Supplement).

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[40] 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[41] to determine the optimal number of clusters, i.e., subphenotypes.

In order to evaluate the reproducibility, we validated our subphenotypes in three 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 (see eMethods in Supplement). Then similar to the primary analysis, we performed agglomerative hierarchical clustering to re-derive subphenotypes and determined optimal cluster number using dendrogram and 'NbClust'. To assess sensitivity to clustering algorithms, we re-derived subphenotypes using the Gaussian mixture model (GMM)[42], 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 eMethods in Supplement).

Second, we used the internal validation cohort and re-derived 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.

Last, 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[43], 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 novel 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.

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[44] based on clinical variables for clustering (showing distributions of subphenotypes within low-dimensional space); 2) chord diagrams[45] showing differences of subphenotypes in terms of abnormal clinical variable groups and comorbidities (see eMethods in Supplement).

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 to assess the association of each SDoH variable with 60-day mortality, adjusting for age and sex. After that, we performed agglomerative hierarchical clustering on the 6 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 to assess the association of SDoH strata with 60-day mortality, adjusting for age and sex, within each subphenotype.

Data availability

All data studied in this work can be downloaded from INSIGHT clinical research network at https://insightcrn.org/our-data/, via request. 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.

Ethical approval and patient consent.

The Institutional Review Board of the Weill Cornell Medicine approved this study (Protocol number: 20-04021948).

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.

Author contributions

ES and FW for conceptualization, investigation, writing, reviewing and editing of the manuscript. CS for investigation, data analysis, drafting, editing and reviewing manuscript. MGW for providing data support, discussion and commenting the manuscript. YZ and JHF for discussion, commenting and editing the manuscript.

Competing interests

The authors have declared that no conflict of interest exists.

References

38. American Community Survey: 2018 Data Release New and Notable.
39. Troyanskaya, O.*, et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520-525 (2001).
40. Murtagh, F. & Legendre, P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification* **31**, 274-295 (2014).
41. Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software* **61**(2014).
42. Reynolds, D.A. Gaussian Mixture Models. *Encyclopedia of biometrics* **741**, 659-663 (2009).
43. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, San Francisco, California, USA, 2016).
44. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
45. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811-2812 (2014).

Tables and Figures

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	8,199	3,519	2,700
Construction method	70% patients (randomly selected) from 4 medical centers of INSIGNT network, including: NYU-LMC, NYP-WCMC, MSHS, and MMC	Remaining 30% patients from 4 medical centers of INSIGNT network, including: NYU-LMC, NYP-WCMC, MSHS, and MMC	NYP-CUMC
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 (%)	3,787 (46.2)	1,585 (45.0)	1,305 (48.3)
Race, N (%)			
White	2,036 (24.8)	838 (23.8)	675 (25.0)
Black	2,155 (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)	-

Abbreviations: ICU = intensive care unit; IQR = Interquartile range; MMC = Montefiore Medical Center; MSHS = Mount Sinai Health System; NYP-CUMC = New York Presbyterian - Columbia University Medical Center; NYP-WCMC = New York Presbyterian - Weill Cornell Medical Center; NYU-LMC = New York University Langone Medical Center; SDoH = social determinants of health.

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)		
ESR, mm/hr, 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	< 0.001
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	0.26
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)	< 0.001	0.04
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)	0.37	0.14
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	< 0.001
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	0.02
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 & Hepatic markers						< 0.001	< 0.001
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)		
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						< 0.001	< 0.001
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)		
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)		
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	< 0.001
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	0.16
Renal markers						< 0.001	< 0.001
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)		
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						< 0.001	< 0.001
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)		
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, $\times 10^3$/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						< 0.001	< 0.001
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)		
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)	0.05	0.06
						< 0.001	0.73
Comorbidity, (missing=590), N (%)							
Hypertension	4744 (62.35)	1238 (49.68)	1696 (60.16)	1095 (78.44)	715 (79.27)		
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	-
						< 0.001	-
Outcomes (60 days), N (%)							
Mortality	1529 (18.65)	188 (6.94)	528 (17.33)	337 (22.68)	476 (49.64)		
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	-
						< 0.001	-
Medications, N (%)							
Antibiotics	2952 (36.00)	731 (27.00)	1219 (40.01)	559 (37.62)	443 (46.19)		
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	-
						< 0.001	-

Abbreviations: 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.

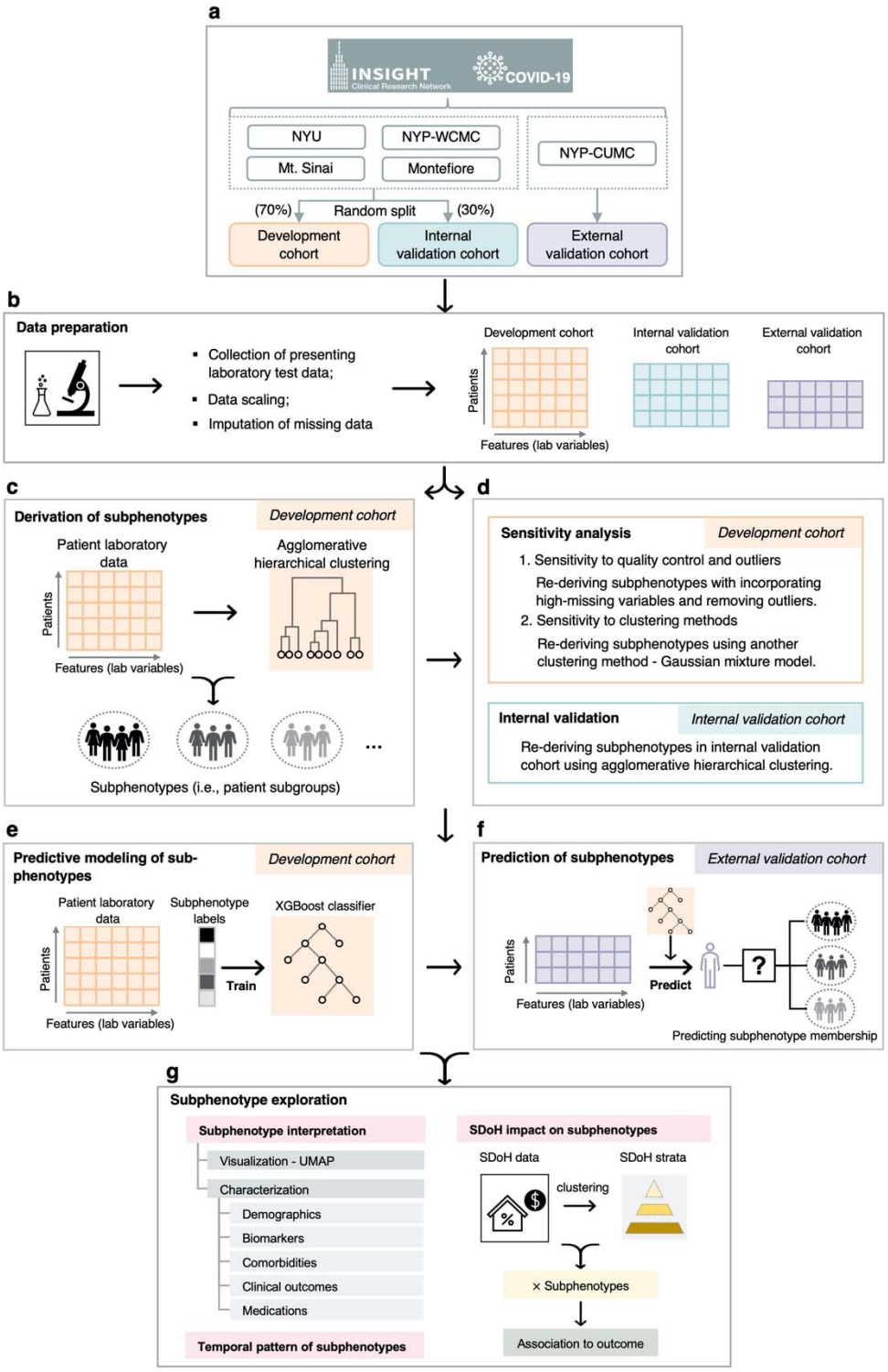

Figure 1. A schematic of the analysis plan.
a. Strategy for construction of development, internal validation, and external validation cohorts. **b.** Data preparation for clustering analysis. **c.** Derivation of subphenotypes in the development

cohort. Reproducibility of the identified subphenotypes were evaluated in multiple ways, including
(d) sensitivity analyses in the development cohort and subphenotype re-derivation 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. **(g)** further analyses were conducted to interpret subphenotypes, explore temporal patterns of subphenotypes during the pandemic, and evaluate impact of SDoH characterisitics on subphenotypes.

Abbreviations: MMC = Montefiore Medical Center; MSHS = Mount Sinai Health System; NYP-CUMC = New York Presbyterian - Columbia University Medical Center; NYP-WCMC = New York Presbyterian - Weill Cornell Medical Center; NYU-LMC = New York University Langone Medical Center; SDoH = social determinants of health; UMAP = Uniform Manifold Approximation and Projection

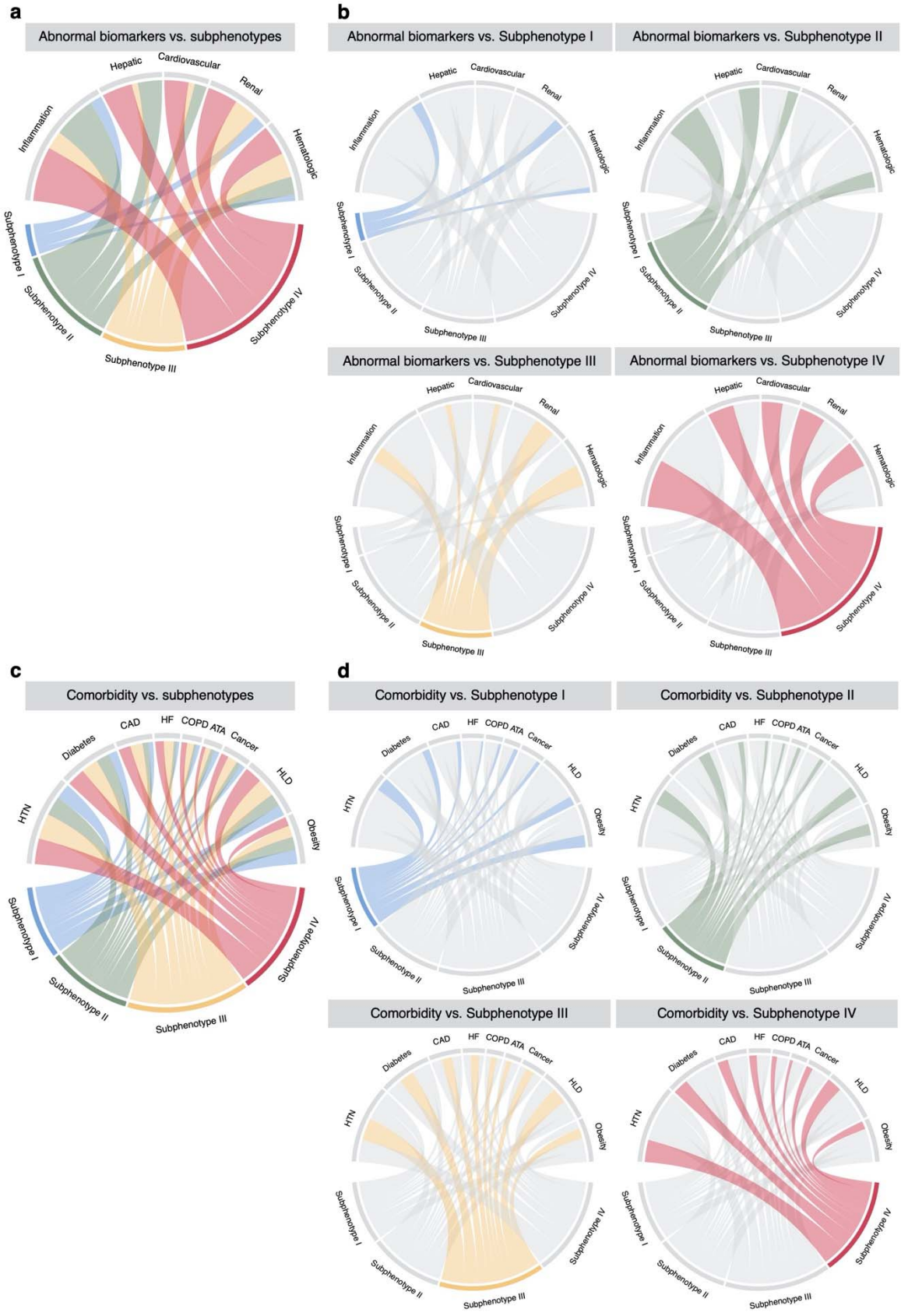
a
Abnormal biomarkers vs. subphenotypes
Hepatic
Cardiovascular
Renal
Hematologic
Subphenotype IV
Subphenotype III
Subphenotype II
Subphenotype I
Inflammation
b
Abnormal biomarkers vs. Subphenotype I
Abnormal biomarkers vs. Subphenotype II
Abnormal biomarkers vs. Subphenotype III
Abnormal biomarkers vs. Subphenotype IV
c
Comorbidity vs. subphenotypes
CAD
HF
COPD
ATA
Cancer
HLD
Obesity
HTN
Diabetes
d
Comorbidity vs. Subphenotype I
Comorbidity vs. Subphenotype II
Comorbidity vs. Subphenotype III
Comorbidity vs. Subphenotype IV

Figure 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

Abbreviations: ATA = asthma; CAD = coronary artery disease; COPD = chronic obstructive pulmonary disease; HF = heart failure; HLD = hyperlipidemia; HTN = hypertension.

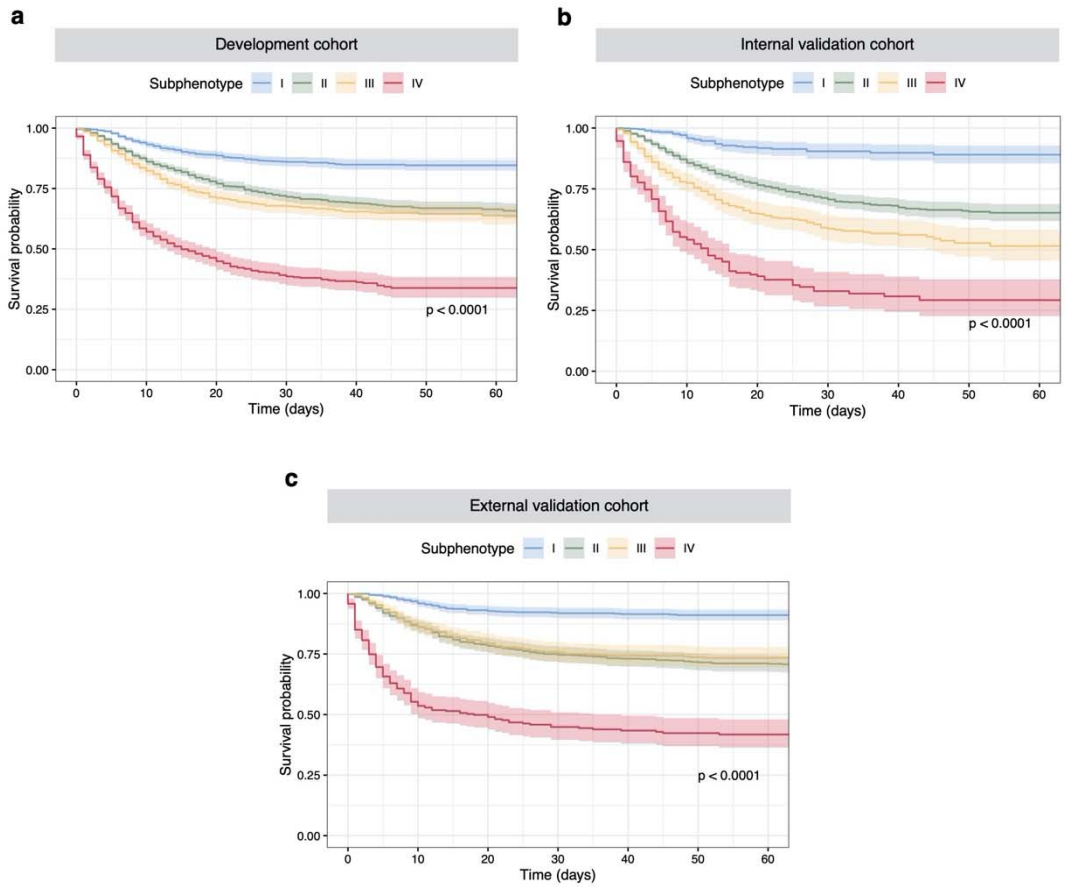

Figure 3. Kaplan-Meier plots for 60-day mortality by subphenotypes. The survival probabilities were shown with 95% confidence interval.

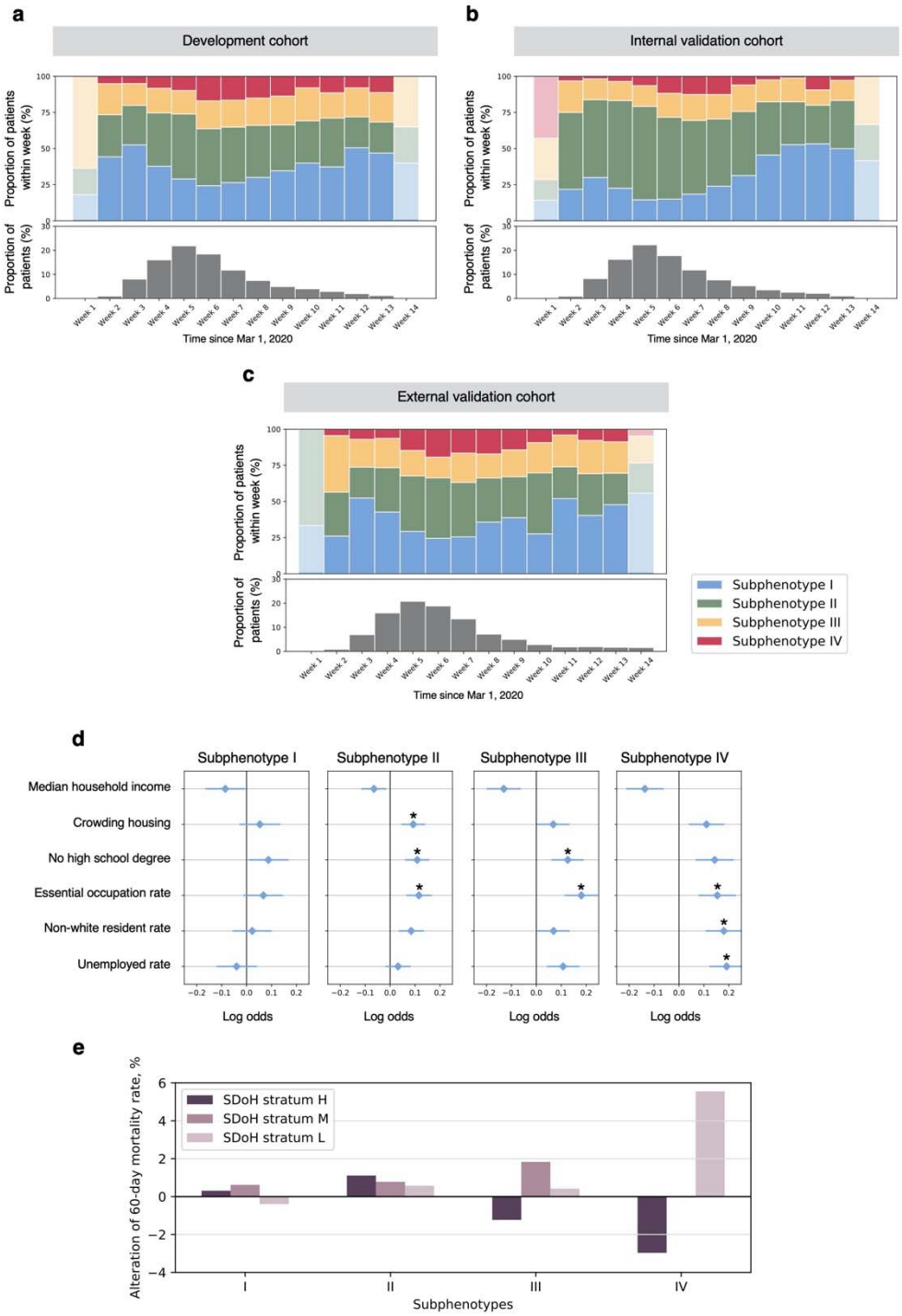

Figure 4. Plots showing temporal patterns and SDoH implications of subphenotypes. **a-c.** Proportions of subphenotype memberships of patients confirmed per week, since March 1, 2020. **d.** Log odds (mean values and standard deviation) showing associations between individual SDoH characteristics and 60-day mortality risk, using logistic regression analysis,

adjusting for age and sex. **e.** Plot showing alteration of 60-day mortality rate of each SDoH stratum to that of subphenotype level.