Title: Clinical and Temporal Characterization of COVID-19 Subgroups Using Patient Vector Embeddings of Electronic Health Records

Authors: Casey N. Ta[1], Jason E. Zucker[2], Po-Hsiang Chiu[1], Yilu Fang[1], Karthik Natarajan[1], Chunhua Weng[1*]

Affiliations:

[1]Department of Biomedical Informatics, Columbia University Irving Medical Center; New York, NY, USA.

[2]Division of Infectious Diseases, Department of Medicine, Columbia University Irving Medical Center; New York, NY, USA.

***Corresponding author**:

Name: Chunhua Weng

Postal address: 622 W. 168th ST, PH-20, New York, NY 10032

Email: chunhua@columbia.edu

Telephone: 212-304-7907

Keywords: COVID-19, SARS-CoV-2, Cluster analysis, Unsupervised machine learning

Word count: 8280

ABSTRACT

Objective: To identify and characterize clinical subgroups of hospitalized COVID-19 patients.

Materials and Methods: Electronic health records of hospitalized COVID-19 patients at NewYork-Presbyterian/Columbia University Irving Medical Center were temporally sequenced and transformed into patient vector representations using Paragraph Vector models. K-means clustering was performed to identify subgroups.

Results: A diverse cohort of 11,313 patients with COVID-19 and hospitalizations between March 2, 2020 and December 1, 2021 were identified; median [IQR] age: 61.2 [40.3-74.3]; 51.5% female. Twenty subgroups of hospitalized COVID-19 patients, labeled by increasing severity, were characterized by their demographics, conditions, outcomes, and severity (mild-moderate/severe/critical). Subgroup temporal patterns were characterized by the durations in each subgroup, transitions between subgroups, and the complete paths throughout the course of hospitalization.

Discussion: Several subgroups had mild-moderate SARS-CoV-2 infections but were hospitalized for underlying conditions (pregnancy, cardiovascular disease (CVD), etc.). Subgroup 7 included solid organ transplant recipients who mostly developed mild-moderate or severe disease. Subgroup 9 had a history of type-2 diabetes, kidney and CVD, and suffered the highest rates of heart failure (45.2%) and end-stage renal disease (80.6%). Subgroup 13 was the oldest (median: 82.7 years) and had mixed severity but high mortality (33.3%). Subgroup 17 had critical disease and the highest mortality (64.6%), with age (median: 68.1 years) being the only notable risk factor. Subgroups 18-20 had critical disease with high complication rates and long hospitalizations (median: 40+ days). All subgroups are detailed in the full text. A chord diagram depicts the most common transitions, and paths with the highest prevalence, longest hospitalizations, lowest and highest mortalities are presented. Understanding these subgroups and their pathways may aid clinicians in their decisions for better management and earlier intervention for patients.

INTRODUCTION

Since the emergence of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) in December 2019, there have been over 400 million confirmed cases of Coronavirus Disease 2019 (COVID-19) worldwide, leading to over 6 million deaths as of April 2022,[1] and over 314,000 hospitalizations in the United States alone.[2] SARS-CoV-2 infections can lead to a wide range of clinical presentations and disease severity, ranging from asymptomatic carriers to critical cases with acute respiratory distress syndrome (ARDS), shock, and multiorgan failure.[3] COVID-19 is also known to manifest heterogeneously, potentially leading to heart failure, renal failure, liver injury, gastrointestinal symptoms, neurological issues, cognitive dysfunction, and systemic manifestations.[4–9]

Electronic health records (EHR) data analyses have been pivotal in contributing to our understanding of COVID-19. Argenziano et al. manually abstracted EHR data of the first 1000 COVID-19 patients admitted at NewYork-Presbyterian/Columbia University Irving Medical Center (NYP/CUIMC) to characterize patient demographics, presenting symptoms and comorbidities, hospital course, and outcomes.[10] Among key observations, Argenziano et al. found high rates of acute renal failure syndrome (ARFS; 78.0% among intensive care unit (ICU) patients) and prolonged intubations (median 28.5 days). Brat et al. and the Consortium for Clinical Characterization of COVID-19 by EHR (4CE) harmonized EHR data of COVID-19 patients across five countries using common data models (CDM) and found that trends towards progressively abnormal values of laboratory measurements for inflammatory, immune, hepatic, coagulatory, and renal function correlated with worsening disease.[11] Weber et al. performed a retrospective cohort study on EHR data using aggregated statistics from 315 hospitals across six countries and found that patients hospitalized in the second wave had a 9.9% reduction in the risk of severe disease relative to the first wave, with significantly lower relative risk among patients aged 26-49 (0.77 [0.63-0.94]) and 50-69 (0.84 [0.72-0.97]), and among Black patients (0.89 [0.81-0.98]).[12]

Kostka et al. conducted an Observational Health Data Sciences and Informatics (OHDSI) Network study to characterize comorbidities, symptoms, medications, and outcomes in three cohorts: COVID-19 patients; hospitalized COVID-19 patients; and hospitalized COVID-19 patients requiring intensive services, finding both similarities and differences in the cohort characteristics across the international sites.[13]

To better understand the heterogeneity of COVID-19, researchers have adopted many approaches to identify and characterize subtypes of the SARS-CoV-2 virus and COVID-19 disease presentation. Morais et al. analyzed global samples of SARS-CoV-2 genomic sequences by identifying segments of high genetic variability and clustering the widely shared polymorphisms, finding six well-defined subtypes with polymorphisms in genes coding for nonstructural, spike, nucleocapsid, and accessory proteins that may confer phenotypic implications.[14] Chen et al. analyzed immune signatures in leukocytes and identified three COVID-19 subtypes that varied by levels of enrichment of immune-inciting and immune-inhibiting signatures; the subtype with the highest immune-inciting and lowest immune-inhibiting signatures had the best outcomes.[15] Huang et al. assessed symptoms at 61+ days post-diagnosis using non-negative matrix factorization on symptom co-occurrences from EHR and found five symptom clusters among long-haulers: chest pain-cough, dyspnea-cough, anxiety-tachycardia, abdominal pain-nausea, and low back pain-joint pain.[16] Lusczek et al. performed ensemble clustering on EHR data collected within 72 hours of hospital admission and found three clinical COVID-19 phenotypes with different comorbidities, complications, and outcomes, including phenotype-III with respiratory comorbidities, phenotype-II patients with moderate severity, and phenotype-I with hematologic, renal, and cardiac comorbidities, and 7.30-fold increase in hazard of death relative to phenotype-III.[17] Sudre et al. applied unsupervised time-series clustering to the first five days of self-reported symptoms, yielding six clusters of symptom presentations, which were predictive of need for respiratory support.[18] In a prospective cohort study, Kenny et al. performed multiple correspondence analysis and hierarchical clustering on self-reported symptoms present more than 4 weeks after symptom onset and found three clusters among long-COVID patients associated predominantly with 1) pain symptoms, 2) cardiovascular symptoms, and 3) fewer long-COVID symptoms.[19] Oh et al. analyzed temporal patterns in EHR data, including lab measurements and treatments, collected during the first 24 hours of ICU admission using sequence clustering methods.[20] Oh et al. identified four clinical subphenotypes in critical COVID-19 patients, ranging from patients with few invasive interventions during the first 24 hours who experienced good outcomes to patients who deteriorated during the first 24 hours and ultimately had poor outcomes.

In this study, we retrospectively analyzed EHR data of hospitalized COVID-19 patients to identify and characterize highly detailed clinical subgroups of COVID-19 patients. We developed an analysis pipeline to extract EHR data into medical coding sequences and transformed them into patient vector representations using Paragraph Vector embedding models. We applied K-means clustering on the patient vectors to identify COVID-19 clinical subgroups and characterized each of the subgroups based on demographics, baseline health prior to SARS-CoV-2 infection, and conditions and outcomes observed. We evaluated the patient vectors daily to characterize temporal patterns between subgroups, identifying subgroup paths with the highest prevalence, longest duration, lowest and highest mortality.

MATERIALS AND METHODS

Figure 1 illustrates the major steps within the study design for clinical and temporal characterization of COVID-19 subgroups using patient vector embeddings of EHR data, including (A) selecting the COVID-19 inpatient cohort and characterization windows, (B) transforming structured EHR data into medical coding sequences (MCS), (C) inferring patient vectors from

MCSs using paragraph vector models, (D) identifying COVID-19 patient clusters, and (E) summarizing clinical and temporal characteristics of each subgroup. The main details are described here with additional details provided in the Supplementary Materials. Supplementary **Table S1** provides a list of abbreviations used throughout the manuscript.

EHR data

We analyzed EHR data from NYP/CUIMC's Observational Medical Outcomes Partnership (OMOP) database. NYP/CUIMC is a quaternary care academic medical center serving New York, NY, and the surrounding area. Longitudinal inpatient and outpatient EHR were collected and stored in the clinical data warehouse (CDW) as part of the routine clinical care. The NYP/CUIMC CDW was converted to OMOP CDM v5.3 on December 5, 2021; the resulting OMOP database contained records spanning from October 1985 to December 2021. Vaccination data for New York City (NYC) residents were also imported from the NYC Vaccine Registry. This study received institutional review board approval with a waiver for informed consent.

Patient vectors

Generating patient vectors consisted of two stages: 1) converting patients' longitudinal EHR data (OMOP CDM) into temporal sequences, and 2) training vector embedding models using the NYP/CUIMC EHR data (Fig. 1).

Medical coding sequences

EHR data for each patient was transformed from the OMOP database into a linear sequence (the MCS) of medical concepts (conditions (e.g., diagnosed diseases, signs or symptoms), drugs, procedures, lab tests, and death) arranged in temporal order. Concepts observed multiple times daily were only recorded in the MCS once per day to minimize the overrepresentation of repetitive data elements. Records observed within the same day were randomly shuffled to dissociate patterns in how the EHR system records and timestamps clinical events (e.g., some events are timestamped at midnight regardless of actual time of occurrence) from affecting the learned vector representation. Consequently, events occurring at the same time (e.g., laboratory test panel measurements) may be distributed randomly throughout a day's sequence, but the relative order of events across different calendar days will be preserved.

Medical coding sequence embedding

We derived patient-centered vector representations by treating each patient's MCS as a document and applying the Paragraph Vector (PV) Distributed Memory (PV-DM) and Distributed Bag-of-Words (PV-DBOW) algorithms.[21] We trained the PV models using data from all patients in the NYP/CUIMC OMOP database and converting patients' entire clinical histories into MCSs. The PV-DM and PV-DBOW models were configured as 100-dimensional vectors and concatenated (200-dimensions total) to become the patient vectors.

COVID-19 clinical subgrouping

COVID-19 inpatient cohort definition

We adapted the cohort definition applied in the OHDSI study.[13] We identified patients with hospitalizations beginning on or after March 1, 2020, with at least one confirmatory diagnosis or at least one positive SARS-CoV-2 diagnostic test result between 21 days before admission to discharge (Fig. 1). The first qualifying event per patient was used. We deviate from the OHDSI definition by removing antibody tests from the inclusion criteria and removing the requirement for a 365-day observation period before hospitalization.

COVID-19 clinical subgroup analysis

To infer vector representations of the COVID-19 patients, we generated MCSs from patients' clinical data starting from admission and including up to the maximum of 1) 28 days following admission (including additional visits) or 2) hospital discharge; henceforth referred to as COVID-characterization window. The COVID-19 MCSs were converted to patient vectors using the PV models described above. To identify COVID-19 clinical subgroups, we applied K-means clustering to the COVID-19 patient vectors. To guide the selection of K (number of clusters), we employed the elbow method (comparing distortion against the number of clusters) and visualized the clustering results using t-distributed stochastic neighbor embedding (t-SNE). We aimed to balance K such that the major visually distinct clusters within the t-SNE plots were well separated while minimizing K to limit the complexity of subgroup comparisons and increase statistical power for detecting differences between subgroups. We evaluated K=[5,10,15,20,25,30,35,40].

COVID-19 clinical subgroup characterization

For each of the identified COVID-19 clinical subgroups, we characterized the group's age, sex, race-ethnicity, hospital start date, hospital length of stay (LOS), primary discharge diagnoses, number of visits within the COVID-characterization window, and number of visits between 730 and 22 days before hospital admission (henceforth referred to as baseline-characterization window). Patients' vaccination statuses prior to hospitalization were evaluated using vaccination records from NYP/CUIMC and the NYC vaccination registry (covering NYC residents). Patients who were not NYC residents and who did not receive their vaccinations from CUIMC may not have their vaccinations captured. Patients were considered fully vaccinated if they received the Janssen, Moderna, or Pfizer-BioNTech vaccines according to protocol along with the minimum delay time prior to their hospitalization. Characteristics of each subgroup were compared against the complementary patients from the COVID-19 cohort. Sex, race-ethnicity, and vaccination rate were evaluated using chi-square, $\alpha=5\times10^{-5}$. Age, hospital start date, hospital LOS, and visit counts were evaluated using the Mann-Whitney U Test, $\alpha=5\times10^{-5}$.

For each of the COVID-19 clinical subgroups, we characterized the subgroup by evaluating the EHR prevalence rates of clinical concepts observed within their COVID-characterization window. Similarly, we evaluated each group's baseline characteristics by evaluating the EHR prevalence rates of condition concepts within their baseline-characterization window. Some concepts known to be associated with COVID-19 were evaluated as concept sets (**Tables S2 and S3**) and compared against the subgroup's complement (chi-square, $\alpha=5\times10^{-6}$). The EHR prevalence rates of all observed conditions with greater than 10% prevalence within the subgroup were individually compared against each subgroup's complement (chi-square, $\alpha=5\times10^{-8}$).

We used a modified version of the 4CE severe COVID-19 phenotyping algorithm[22] to label patients with non-severe, severe, or critical COVID-19. Patients were labeled as "critical" if the patients had at least one record of death, ARDS, septic shock, or invasive mechanical ventilation (**Table S4**). Patients without the above records but with at least one record of acute hypoxemic respiratory failure (AHRF), hypoxemia, or noninvasive oxygen therapy were labeled "severe". All other patients were labeled "mild-moderate". The frequency of the severity phenotypes in each subgroup was compared to the subgroup's complement (chi-square, $\alpha=5\times10^{-6}$).

By default, K-means clusters are identified by random numeric labels. We relabeled each subgroup in ascending order of mean severity by assigning scores of 1, 2, and 3 for severity phenotypes mild-moderate, severe, and critical, respectively.

COVID-19 clinical subgroup generalizability

Prior to performing the subgroup analysis, 20% of the COVID-19 inpatient cohort were held out to evaluate generalizability of the subgroups. 10% of the patients with the most recent hospitalizations were held out for out-of-time generalizability evaluation, and 10% of patients were randomly selected from the remaining patients for in-time generalizability evaluation. The remaining 80% of patients were used in the analyses above. For each subgroup, we calculated the 95th-percentile distance of training patient vectors to their clusters' centroids. For a given holdout patient vector, its subgroup membership can be predicted via minimum Euclidean distance to the centroids of the trained K-means model. We then calculated the distance of the holdout patient vectors to their predicted clusters' centroids, counted the number of holdout patient vectors farther than the 95th-percentile distance among the training patient vectors, and compared this to the expected frequency of 5% (chi-square, α=0.05).

COVID-19 subgroup temporal analysis

For each patient whose primary discharge diagnoses included COVID-19, sepsis, or viral pneumonia, we used the methods described above to create patient vectors for each day of hospitalization, containing cumulative EHR data from admission up to the given date, and assigned the nearest COVID-19 subgroup. To characterize the temporal relationships between subgroups, we counted the transitions between subgroups and the durations in each subgroup. We identified distinct subgroup paths as the observed series of contiguous COVID subgroups that the patients transitioned between during hospitalization. We grouped discharge statuses into 1) death or discharge to hospice; 2) discharge to additional care services; or 3) discharge (**Table S5**). We counted the number of patients following each path and determined the median total duration.

To determine if any of the subgroups are associated with prognosis, we compared the frequency of death or discharge to hospice among patients who began their hospitalization in each of the subgroups to the rate among the subgroup's complement (chi-square, α=0.0025).

Statistical analyses

The statistical tests (two-sided) and significance levels are described alongside each evaluation. We used the Python packages SciPy v1.4.1 to perform chi-square and Mann-Whitney U tests, SciKit-Learn v0.22.2 for K-means and t-SNE, and Gensim v3.8.1 for paragraph vector embedding.[23]

RESULTS

Hospitalized COVID-19 cohort characteristics

11,313 patients were identified with hospitalizations and either a condition code for COVID-19 or a positive test for SARS-CoV-2. Hospitalization start dates ranged from March 2, 2020 to December 1, 2021. **Figure 2** shows the distribution of a) age, b) race-ethnicity, c) hospitalization LOS, d) hospitalization start date, e) number of prior healthcare visits within the baseline-characterization window, and f) number of healthcare visits within the COVID-characterization window for the full inpatient COVID-19 cohort. The median [interquartile range (IQR)] age was 61.2 [40.3-74.3]. There were four modes in the age distribution, with peaks occurring in the 0-4, 30-34, 65-69, and 80+ age groups. 51.5% of patients were female. Race-ethnicity included 40.1% Hispanic, 3.2% Asian, American Indian and Alaska Native, and Native Hawaiian and Pacific Islander, 10.5% Black, 32.7% White, and 13.4% other/unknown. Most hospitalizations were between 3-9 days. There were three peaks in the hospitalization dates, largely corresponding with the reported number of new cases in NYC: April 2020, January 2021, and August 2021. Many of the patients hospitalized with COVID-19 had very little healthcare utilization at NYP/CUIMC in the prior 2 years, with 30.5% and 12.6% of patients having 0 and 1-2 visits, respectively. However, 21.9% of patients had high healthcare utilization, with 20 or more visits. Within the COVID-characterization window, for 51.2% of patients, the hospitalization was their only visit. 5.4% of patients were fully vaccinated prior to their hospitalization.

COVID-19 clinical subgroup analysis

10% (1131/11313) of patients with the most recent hospitalizations (July 20, 2021 – December 1, 2021) were held out for the out-of-time generalizability evaluation, and 10% (1131/11313) were randomly held out for the in-time generalizability evaluation. 80% (9051/11313) of patients were used for the subgroup analysis.

For K-means clustering, we selected K=20 based on subjective analysis using the elbow method and t-distributed stochastic neighbor embedding (t-SNE) scatterplot. **Figure 3** shows the t-SNE scatterplot of COVID-19 patient vectors. The subgroup labels (Subgroups 1-20) were assigned in ascending order of each subgroup's severity score, e.g., Subgroup 1 and Subgroup 20 identify the subgroups with the lowest and highest severity scores, respectively. Subgroup labels will be abbreviated from Subgroup # to SG#

COVID-19 clinical subgroup characteristics

Table 1 shows the summary statistics of the full COVID-19 cohort and the clinical subgroups. Subgroups 2, 6, and 8 were the most common, comprising 16.9%, 13.5%, and 9.2% of the cohort, respectively. Subgroups 4 and 13 had the youngest and oldest populations, with median ages of 6.9 [1.9-14.1] and 82.7 [74.0-89.4], respectively. SG1 was comprised almost entirely of female patients (98.8%), while SG19 had the highest percentage of male patients (68.8%). SG3 had the highest vaccination rate (8.9%). SG2 had the shortest median hospital duration (3 days), while SG18 had the longest (46 days). The median [IQR] number of visits within the COVID-characterization window was 1 [1-3] visits and did not vary widely among subgroups, although the subgroups with lower severities tended to have more visits. The number of visits within the baseline-characterization window had a larger spread: most patients in Subgroups 15, 18, and 20 had no prior visits, whereas SG7 had a median of 58 visits.

Table 2 shows the prevalence rates of selected clinical concepts within the EHR within the COVID-characterization window for each subgroup compared to each subgroup's complement (chi-square; $\alpha=5\times10^{-6}$). **Table S6** shows the prevalence rates within the COVID-characterization window in each subgroup of individual conditions with at least 10% subgroup prevalence and significant difference with the subgroup's complement (chi-square; $\alpha=5\times10^{-8}$). **Table 3** and **Table S7** show the analogous results from the baseline-characterization window. **Table S8** shows the three most common primary discharge diagnoses (PDD) for each subgroup. Detailed discussion of these results per COVID-19 clinical subgroup will follow in the Discussion section.

Table 1. Summary statistics of the full COVID-19 cohort and subgroups. Values were compared to each subgroup's complement; bold: significant difference; blue/red: subgroup values were significantly less/greater than the complementary cohort values. Sex, race-ethnicity, and vaccinated: count (percentage); chi-square, α=5×10^{-5}. Age, hospital start date, hospital length, number of visits, number of prior visits: median [interquartile range]; Mann-Whitney U Test, α=5×10^{-5}. AIAN: American Indian and Alaska Native; NHPI: Native Hawaiian and Pacific Islander.

	Full	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Count	9051 (100%)	740 (8.2%)	1530 (16.9%)	372 (4.1%)	274 (3.0%)	176 (1.9%)	1225 (13.5%)	242 (2.7%)	831 (9.2%)	263 (2.9%)	263 (2.9%)
Age	61.6 [41.7, 74.4]	**31.3** **[26.9, 35.8]**	**56.7** **[34.9, 72.0]**	60.3 [47.0, 69.0]	**6.9** **[1.9, 14.1]**	65.5 [56.3, 72.8]	**57.4** **[43.2, 69.2]**	62.1 [50.9, 68.7]	**68.3** **[57.9, 77.9]**	**64.6** **[56.0, 75.2]**	61.0 [48.4, 72.3]
Sex (female)	4599 (50.8%)	**731 (98.8%)**	801 (52.4%)	**186 (50.0%)**	**133 (48.5%)**	**60 (34.1%)**	650 (53.1%)	**89 (36.8%)**	**403 (48.5%)**	**88 (33.5%)**	**115 (43.7%)**
Race-ethnicity											
Hispanic	2686 (29.7%)	**297 (40.1%)**	**307 (20.1%)**	**33 (8.9%)**	**50 (18.2%)**	**18 (10.2%)**	**440 (35.9%)**	**48 (19.8%)**	**325 (39.1%)**	**98 (37.3%)**	77 (29.3%)
Asian,AIAN,NHPI	190 (2.1%)	**24 (3.2%)**	**44 (2.9%)**	**6 (1.6%)**	**8 (2.9%)**	**4 (2.3%)**	**23 (1.9%)**	**11 (4.5%)**	**8 (1.0%)**	**7 (2.7%)**	6 (2.3%)
Black	1527 (16.9%)	**78 (10.5%)**	**234 (15.3%)**	**44 (11.8%)**	**44 (16.1%)**	**19 (10.8%)**	**225 (18.4%)**	**54 (22.3%)**	**151 (18.2%)**	**82 (31.2%)**	55 (20.9%)
White	2857 (31.6%)	**242 (32.7%)**	**617 (40.3%)**	**212 (57.0%)**	**118 (43.1%)**	**83 (47.2%)**	**285 (23.3%)**	**104 (43.0%)**	**188 (22.6%)**	**45 (17.1%)**	75 (28.5%)
Other,Unknown	1791 (19.8%)	**99 (13.4%)**	**328 (21.4%)**	**77 (20.7%)**	**54 (19.7%)**	**52 (29.5%)**	**252 (20.6%)**	**25 (10.3%)**	**159 (19.1%)**	**31 (11.8%)**	50 (19.0%)
Vaccinated (%)	227 (2.5%)	23 (3.1%)	**73 (4.8%)**	**33 (8.9%)**	**0 (0.0%)**	6 (3.4%)	**14 (1.1%)**	**21 (8.7%)**	**8 (1.0%)**	**6 (2.3%)**	**5 (1.9%)**
Hospital start date	2020-10-29 [2020-04-18, 2021-02-04]	**2020-11-03** **[2020-06-03, 2021-03-08]**	**2020-11-10** **[2020-09-07, 2021-02-04]**	**2020-11-23** **[2020-09-15, 2021-03-05]**	2020-08-19 [2020-05-09, 2021-01-09]	**2020-11-30** **[2020-09-29, 2021-02-01]**	2020-12-19 [2020-04-13, 2021-02-12]	2020-11-05 [2020-05-09, 2021-02-18]	2020-12-21 [2020-04-17, 2021-02-13]	2020-07-13 [2020-04-10, 2021-01-22]	2020-11-05 [2020-06-01, 2021-02-15]
Hospital length (days)	6.0 [3.0, 11.0]	**4.0** **[3.0, 4.0]**	**3.0** **[2.0, 4.0]**	**4.0** **[2.0, 6.0]**	5.0 [4.0, 10.0]	**10.0** **[7.0, 17.0]**	**4.0** **[3.0, 6.0]**	**9.0** **[7.0, 18.0]**	**5.0** **[4.0, 7.0]**	**9.0** **[6.0, 14.0]**	**12.0** **[7.0, 21.5]**
Number of visits	1.0 [1.0, 3.0]	**2.0** **[2.0, 3.0]**	**2.0** **[1.0, 3.0]**	**2.0** **[2.0, 3.0]**	**3.0** **[1.0, 4.0]**	2.0 [1.0, 3.0]	**1.0** **[1.0, 2.0]**	**3.0** **[1.0, 5.0]**	**1.0** **[1.0, 2.0]**	1.0 [1.0, 2.0]	**2.0** **[1.0, 4.0]**
Number of prior visits	4.0 [0.0, 17.0]	**15.0** **[9.0, 22.0]**	**7.0** **[2.0, 19.0]**	**7.0** **[3.0, 17.0]**	4.0 [0.0, 25.8]	5.0 [1.0, 14.0]	**1.0** **[0.0, 8.0]**	**58.0** **[34.2, 86.0]**	**2.0** **[0.0, 11.0]**	**15.0** **[2.0, 39.0]**	**12.0** **[2.0, 42.0]**

	Full	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Count	9051 (100%)	349 (3.9%)	650 (7.2%)	720 (8.0%)	454 (5.0%)	43 (0.5%)	58 (0.6%)	367 (4.1%)	220 (2.4%)	160 (1.8%)	114 (1.3%)
Age	61.6 [41.7, 74.4]	**71.8** **[59.0, 84.2]**	**63.3** **[51.0, 75.0]**	**82.7** **[74.0, 89.4]**	**70.2** **[60.8, 78.3]**	**9.7** **[4.2, 15.7]**	68.3 [58.3, 75.4]	**68.1** **[57.9, 76.8]**	60.5 [50.6, 70.6]	66.0 [57.2, 73.4]	61.5 [51.3, 70.0]
Sex (female)	4599 (50.8%)	**163 (46.7%)**	**294 (45.2%)**	**352 (48.9%)**	**173 (38.1%)**	**16 (37.2%)**	**20 (34.5%)**	**149 (40.6%)**	**86 (39.1%)**	**50 (31.2%)**	**40 (35.1%)**
Race-ethnicity											
Hispanic	2686 (29.7%)	**29 (8.3%)**	**241 (37.1%)**	223 (31.0%)	**193 (42.5%)**	7 (16.3%)	7 (12.1%)	125 (34.1%)	82 (37.3%)	46 (28.7%)	40 (35.1%)
Asian,AIAN,NHPI	190 (2.1%)	**15 (4.3%)**	**9 (1.4%)**	8 (1.1%)	**2 (0.4%)**	0 (0.0%)	3 (5.2%)	6 (1.6%)	3 (1.4%)	0 (0.0%)	3 (2.6%)
Black	1527 (16.9%)	**52 (14.9%)**	**113 (17.4%)**	127 (17.6%)	**77 (17.0%)**	5 (11.6%)	9 (15.5%)	73 (19.9%)	29 (13.2%)	34 (21.2%)	22 (19.3%)
White	2857 (31.6%)	**187 (53.6%)**	**145 (22.3%)**	224 (31.1%)	**82 (18.1%)**	18 (41.9%)	30 (51.7%)	86 (23.4%)	48 (21.8%)	38 (23.8%)	30 (26.3%)
Other,Unknown	1791 (19.8%)	**66 (18.9%)**	**142 (21.8%)**	138 (19.2%)	**100 (22.0%)**	13 (30.2%)	9 (15.5%)	77 (21.0%)	58 (26.4%)	42 (26.2%)	19 (16.7%)
Vaccinated (%)	227 (2.5%)	18 (5.2%)	**1 (0.2%)**	**9 (1.2%)**	**1 (0.2%)**	**0 (0.0%)**	**1 (1.7%)**	**2 (0.5%)**	**2 (0.9%)**	**4 (2.5%)**	**0 (0.0%)**
Hospital start date	2020-10-29 [2020-04-18, 2021-02-04]	**2021-01-26** **[2020-12-16, 2021-03-06]**	**2020-04-08** **[2020-03-30, 2020-04-30]**	**2020-05-03** **[2020-04-08, 2021-01-10]**	**2020-04-30** **[2020-04-09, 2021-01-13]**	2020-04-30 [2020-04-08, 2021-01-29]	**2021-01-24** **[2020-12-06, 2021-02-26]**	**2020-04-16** **[2020-04-01, 2020-12-29]**	2020-12-03 [2020-04-16, 2021-02-09]	**2020-04-21** **[2020-04-03, 2021-01-08]**	**2020-03-31** **[2020-03-26, 2020-04-06]**
Hospital length (days)	6.0 [3.0, 11.0]	**9.0** **[6.0, 15.0]**	**9.0** **[6.0, 13.0]**	**7.0** **[4.0, 10.0]**	**12.0** **[8.0, 17.0]**	**45.0** **[22.5, 81.0]**	**41.0** **[20.0, 62.8]**	**13.0** **[7.0, 19.5]**	**46.0** **[32.8, 69.0]**	**40.0** **[26.0, 59.2]**	**45.0** **[30.0, 61.0]**
Number of visits	1.0 [1.0, 3.0]	1.0 [1.0, 2.0]	**1.0** **[1.0, 2.0]**	**1.0** **[1.0, 2.0]**	**1.0** **[1.0, 2.0]**	1.0 [1.0, 2.0]	**1.0** **[1.0, 1.0]**	**1.0** **[1.0, 1.0]**	**1.0** **[1.0, 1.0]**	**1.0** **[1.0, 1.0]**	**1.0** **[1.0, 2.0]**
Number of prior visits	4.0 [0.0, 17.0]	**3.0** **[0.0, 10.0]**	**1.0** **[0.0, 10.0]**	**3.0** **[0.0, 12.0]**	**2.0** **[0.0, 13.0]**	0.0 [0.0, 12.0]	2.0 [0.0, 9.0]	**1.0** **[0.0, 10.0]**	**0.0** **[0.0, 8.0]**	**1.0** **[0.0, 5.0]**	**0.0** **[0.0, 4.8]**

Table 2. **COVID-19 subgroup concept prevalence rates.** The prevalence rates during the COVID-characterization window of relevant concept sets and COVID-19 severity are shown for each subgroup and compared to each subgroup's complement (chi-square; α=5×10^{-6}). Bold: significant difference; blue/red: subgroup rates were significantly less/greater than the complementary cohort rates. AHRF: acute hypoxemic respiratory failure; ARDS: acute respiratory distress syndrome; ARFS: Acute renal failure syndrome; ESRD: end-stage renal disease; Altered mental S: Altered mental status; Ventilation: invasive respiratory ventilation.

	Full	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Count	9051	740	1530	372	274	176	1225	242	831	263	263	349	650	720	454	43	58	367	220	160	114
Fever	33.5%	**5.3%**	**5.4%**	**4.0%**	**60.2%**	**10.2%**	**49.1%**	36.0%	34.1%	36.5%	33.8%	35.8%	**62.8%**	36.4%	**52.2%**	**69.8%**	53.4%	42.8%	**59.5%**	**53.1%**	**78.9%**
Cough	27.3%	**2.8%**	**3.3%**	**2.4%**	**12.8%**	**5.7%**	**52.5%**	22.7%	**37.3%**	27.4%	**11.8%**	30.7%	**59.2%**	26.1%	**46.5%**	16.3%	25.9%	34.9%	33.2%	36.2%	**56.1%**
Diarrhea	10.0%	**0.9%**	**2.9%**	**0.5%**	16.4%	0.6%	**14.7%**	**27.3%**	11.0%	10.6%	10.6%	16.6%	**16.0%**	5.4%	13.9%	20.9%	20.7%	10.4%	17.3%	16.9%	17.5%
Constipation	11.3%	**4.1%**	**5.0%**	7.5%	11.7%	14.2%	**6.0%**	9.9%	8.5%	11.8%	**25.9%**	**20.6%**	9.4%	**16.9%**	17.6%	25.6%	27.6%	12.3%	**28.2%**	**35.0%**	**29.8%**
Dyspnea	44.6%	**3.4%**	**7.0%**	**3.8%**	**11.3%**	34.7%	**68.5%**	52.5%	**56.4%**	51.0%	36.1%	**63.0%**	**80.3%**	43.5%	**76.2%**	23.3%	58.6%	**74.9%**	**80.9%**	**86.9%**	**86.0%**
Viral pneumonia	48.0%	**3.8%**	**2.8%**	**0.8%**	**11.7%**	**5.1%**	**69.6%**	47.5%	**66.1%**	56.3%	**24.7%**	**71.6%**	**91.7%**	**62.4%**	**90.5%**	32.6%	74.1%	**83.1%**	**83.2%**	**85.0%**	**95.6%**
Hypoxemia	38.6%	**1.2%**	**2.5%**	**5.1%**	**15.7%**	31.2%	**47.2%**	42.6%	44.6%	38.4%	28.5%	**64.2%**	**68.3%**	45.0%	**70.9%**	25.6%	**69.0%**	**79.3%**	**95.0%**	**86.2%**	**85.1%**
AHRF	35.3%	**0.3%**	**1.5%**	**1.3%**	**14.2%**	**14.2%**	35.3%	27.7%	38.4%	37.3%	**18.3%**	**47.3%**	**67.8%**	**45.8%**	**74.2%**	65.1%	**79.3%**	**86.1%**	**95.9%**	**94.4%**	**97.4%**
ARDS	13.5%	**0.3%**	**1.0%**	**0.5%**	19.0%	5.7%	**3.1%**	8.7%	**5.2%**	5.3%	5.7%	5.4%	9.5%	11.4%	14.5%	**65.1%**	**70.7%**	**68.1%**	**92.3%**	**87.5%**	**100.0%**
Pleural effusion	15.2%	**0.1%**	**2.5%**	**4.6%**	14.2%	**71.0%**	**5.3%**	**28.5%**	**5.9%**	22.1%	**35.0%**	**26.1%**	**6.6%**	12.6%	10.1%	**67.4%**	**82.8%**	**41.4%**	**77.3%**	**65.6%**	**43.0%**
Atelectasis	11.1%	**1.4%**	**3.7%**	7.0%	16.8%	**56.2%**	7.3%	15.7%	6.7%	9.9%	**29.3%**	18.1%	6.2%	11.2%	6.2%	**76.7%**	**50.0%**	14.7%	**33.2%**	**36.9%**	18.4%
Anemia	12.4%	**5.5%**	**3.8%**	**3.8%**	9.9%	23.3%	**6.4%**	17.8%	10.0%	19.0%	**30.8%**	18.1%	**6.3%**	14.3%	10.6%	34.9%	**46.6%**	19.9%	**47.7%**	**49.4%**	**43.9%**
Tachycardia	23.6%	**4.9%**	**10.0%**	**10.8%**	20.8%	**44.3%**	20.8%	31.0%	22.6%	21.3%	**39.5%**	21.5%	21.8%	26.7%	31.7%	**65.1%**	44.8%	**41.7%**	**71.8%**	**65.6%**	**64.9%**
Heart failure	13.1%	**0.1%**	**6.3%**	**1.1%**	4.4%	**35.2%**	**6.0%**	15.3%	15.5%	**45.2%**	14.1%	20.6%	8.3%	**24.2%**	15.9%	**44.2%**	**36.2%**	**24.0%**	22.3%	**31.2%**	16.7%
ARFS	30.3%	**1.5%**	**4.7%**	**4.8%**	**6.9%**	30.7%	**15.8%**	**63.6%**	33.5%	38.8%	**47.1%**	39.0%	28.6%	**52.9%**	**52.2%**	44.2%	**81.0%**	**75.2%**	**89.5%**	**85.6%**	**88.6%**
ESRD	5.3%	**0.3%**	**1.8%**	0.3%	0.0%	4.5%	**1.1%**	**30.6%**	**0.5%**	**80.6%**	3.0%	4.6%	**0.3%**	**1.0%**	**0.2%**	2.3%	**25.9%**	6.0%	**15.0%**	**15.0%**	11.4%
Sepsis	14.4%	**0.1%**	**0.7%**	**0.5%**	5.8%	8.0%	**5.1%**	18.2%	**5.5%**	19.0%	**26.6%**	12.0%	9.7%	**22.1%**	14.5%	27.9%	**60.3%**	**56.9%**	**80.0%**	**81.9%**	**80.7%**
Septic shock	8.5%	**0.0%**	**0.2%**	**0.0%**	2.9%	5.1%	**0.3%**	5.0%	**0.6%**	5.7%	13.3%	4.6%	**0.9%**	6.9%	2.6%	23.3%	**56.9%**	**50.1%**	**79.5%**	**71.9%**	**70.2%**
Delirium	6.0%	**0.0%**	**0.1%**	1.6%	0.7%	4.0%	**0.7%**	6.6%	3.6%	6.1%	11.8%	7.7%	3.7%	**19.7%**	9.0%	16.3%	19.0%	9.8%	**26.8%**	**32.5%**	**25.4%**
Altered mental S	11.2%	**0.3%**	**1.4%**	4.3%	4.7%	2.8%	**1.7%**	9.9%	8.8%	**23.2%**	**24.3%**	17.2%	**3.4%**	**33.8%**	13.7%	20.9%	**31.0%**	**25.1%**	**42.3%**	**39.4%**	**46.5%**
Ventilation	11.4%	**0.1%**	**0.3%**	**0.5%**	10.2%	14.8%	**0.2%**	10.3%	**0.8%**	**2.3%**	9.5%	3.7%	**1.1%**	**4.0%**	**1.1%**	**65.1%**	**82.8%**	**80.4%**	**96.4%**	**98.8%**	**100.0%**
Dead	11.6%	**0.0%**	**1.3%**	**0.3%**	**0.4%**	4.0%	**1.8%**	6.6%	**6.4%**	11.4%	17.5%	10.3%	9.8%	**33.3%**	13.4%	11.6%	**41.4%**	**64.6%**	**45.0%**	**38.1%**	**26.3%**
Mild-Moderate	50.2%	**98.2%**	**95.3%**	**92.7%**	**66.4%**	56.8%	**40.4%**	51.2%	**43.2%**	45.2%	54.0%	**23.8%**	**15.4%**	**34.7%**	**12.1%**	**11.6%**	**3.4%**	**0.3%**	**0.5%**	**0.0%**	**0.0%**
Severe	30.2%	**1.4%**	**3.2%**	**6.5%**	**19.7%**	25.0%	**56.7%**	34.7%	**49.6%**	38.4%	18.6%	**61.3%**	**71.7%**	**27.4%**	**66.7%**	**16.3%**	**3.4%**	**5.2%**	**0.0%**	**0.6%**	**0.0%**
Critical	19.6%	**0.4%**	**1.5%**	**0.8%**	**13.9%**	18.2%	**2.9%**	14.0%	**7.2%**	16.3%	27.4%	**14.9%**	**12.9%**	**37.9%**	**21.1%**	**72.1%**	**93.1%**	**94.6%**	**99.5%**	**99.4%**	**100.0%**

Table 3. **Subgroup condition prevalence rates at baseline.** The prevalence rates within the baseline-characterization window of relevant concept sets are shown for each COVID-19 subgroup and compared to each subgroup's complement (chi-square; α=5×10^{-6}). Bold: significant difference; blue/red: subgroup rates were significantly less/greater than the complementary cohort rates. T2DM: type-2 diabetes mellitus; COPD: chronic obstructive pulmonary disease; Atherosclerosis CA: atherosclerosis of coronary artery; ARFS: acute renal failure syndrome; CKD: chronic kidney disease; ESRD: end-stage renal disease; SOT: solid organ transplant.

	Full	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Count	9051	740	1530	372	274	176	1225	242	831	263	263	349	650	720	454	43	58	367	220	160	114
Obesity	12.9%	**28.9%**	15.8%	8.1%	6.9%	12.5%	9.8%	17.4%	14.2%	22.1%	16.7%	6.3%	8.2%	**6.4%**	12.1%	2.3%	8.6%	11.2%	7.3%	6.9%	8.8%
Hypertension	31.5%	**4.9%**	31.8%	**20.4%**	**7.7%**	36.4%	**24.0%**	**84.7%**	**43.3%**	**57.0%**	42.6%	28.7%	26.5%	**43.9%**	40.7%	11.6%	36.2%	32.2%	27.7%	24.4%	24.6%
Hyperlipidemia	21.8%	**0.7%**	22.0%	14.0%	**1.1%**	28.4%	**15.0%**	**59.5%**	**31.5%**	**54.4%**	24.0%	18.1%	15.4%	**31.4%**	**33.7%**	0.0%	24.1%	24.8%	19.1%	16.9%	14.9%
T2DM	19.1%	**1.4%**	**14.7%**	**6.5%**	**2.2%**	19.3%	**5.4%**	**53.3%**	**47.4%**	**57.8%**	21.3%	**9.2%**	**6.5%**	22.2%	**43.8%**	2.3%	20.7%	25.9%	21.4%	18.8%	15.8%
Pregnant	8.2%	**88.6%**	**1.8%**	**1.1%**	4.0%	0.0%	**1.4%**	**0.0%**	**0.4%**	0.8%	0.4%	**0.6%**	**0.9%**	**0.0%**	**0.7%**	2.3%	0.0%	**1.4%**	0.0%	0.0%	0.0%
Constipation	9.7%	**4.6%**	8.0%	10.2%	**18.2%**	2.8%	6.2%	**23.6%**	9.3%	**23.2%**	**26.6%**	6.0%	7.2%	**15.1%**	9.0%	4.7%	8.6%	7.9%	8.6%	4.4%	5.3%
Diarrhea	5.0%	2.3%	4.4%	4.0%	**12.0%**	4.0%	3.7%	**24.8%**	4.8%	9.1%	**12.9%**	5.7%	3.4%	3.5%	3.1%	7.0%	8.6%	4.1%	3.6%	1.2%	0.9%
Cough	10.0%	7.8%	8.2%	4.8%	14.2%	9.1%	11.0%	**35.1%**	9.1%	17.9%	10.6%	7.4%	9.2%	11.1%	8.8%	4.7%	3.4%	10.1%	7.3%	7.5%	4.4%
Dyspnea	15.7%	**7.2%**	17.5%	9.9%	10.6%	27.8%	13.1%	**52.5%**	13.7%	**44.9%**	24.0%	9.5%	12.9%	14.2%	16.5%	4.7%	13.8%	14.4%	11.8%	10.0%	4.4%
COPD	4.5%	**0.3%**	3.7%	3.0%	1.1%	5.1%	5.0%	**10.7%**	7.0%	9.5%	4.9%	2.3%	4.9%	5.3%	5.1%	0.0%	1.7%	7.9%	3.6%	0.6%	0.9%
Pleural effusion	4.3%	**0.4%**	3.1%	3.2%	4.7%	8.0%	1.9%	**21.1%**	2.5%	**16.0%**	**13.7%**	4.3%	2.6%	4.4%	2.6%	4.7%	5.2%	6.8%	3.6%	5.0%	1.8%
Atherosclerosis CA	12.1%	**0.1%**	13.0%	6.2%	**0.4%**	**28.4%**	**5.3%**	**32.6%**	15.8%	**44.1%**	13.3%	12.6%	6.6%	**18.6%**	16.3%	2.3%	17.2%	14.4%	8.2%	10.0%	2.6%
Heart disease	21.3%	**1.4%**	24.4%	**10.8%**	13.5%	**55.7%**	**10.8%**	**55.0%**	23.6%	**61.6%**	28.9%	24.1%	**13.4%**	**30.8%**	23.6%	16.3%	25.9%	24.0%	13.2%	15.6%	7.0%
ARFS	9.6%	**0.4%**	7.0%	4.6%	6.6%	9.1%	**3.8%**	**40.5%**	10.6%	**34.6%**	**21.3%**	8.9%	4.6%	**15.6%**	12.3%	7.0%	10.3%	13.4%	9.5%	7.5%	4.4%
CKD	12.0%	**0.1%**	**8.4%**	**3.0%**	3.6%	13.1%	**4.2%**	**67.8%**	13.2%	**73.4%**	12.9%	11.2%	**4.9%**	16.0%	14.5%	0.0%	12.1%	14.7%	11.8%	10.6%	7.9%
ESRD	4.4%	**0.1%**	2.9%	0.8%	1.5%	4.0%	**1.1%**	**34.7%**	**0.7%**	**59.3%**	2.3%	3.4%	0.9%	1.5%	0.9%	0.0%	6.9%	4.1%	5.0%	3.1%	1.8%
Vitamin D deficiency	5.0%	1.6%	5.2%	5.4%	5.5%	4.5%	3.4%	**26.0%**	4.6%	**14.1%**	6.1%	6.0%	2.3%	4.7%	3.1%	4.7%	1.7%	3.5%	4.5%	3.8%	2.6%
Immunodeficiency	1.9%	0.1%	1.4%	1.3%	5.5%	0.6%	0.6%	**29.8%**	0.6%	0.8%	3.0%	4.0%	0.3%	0.1%	0.0%	0.0%	8.6%	1.6%	1.4%	0.0%	0.0%
Dementia	4.3%	**0.0%**	**1.6%**	0.5%	0.0%	0.0%	**1.7%**	0.4%	4.6%	9.1%	0.8%	6.0%	2.3%	**25.3%**	7.3%	0.0%	0.0%	5.2%	1.8%	1.2%	2.6%
SOT	4.4%	**0.4%**	3.3%	1.6%	5.8%	2.3%	**0.9%**	**86.4%**	1.7%	7.2%	2.7%	2.9%	0.8%	**0.4%**	0.2%	11.6%	5.2%	2.5%	3.6%	4.4%	4.4%

Table 4. Holdout evaluation. The K-means model trained on the training set was used to predict the COVID-19 clinical subgroup of each holdout patient vector. We counted the number of holdout patient vectors that were farther than the 95th percentile distance of the training patient vectors from their respective cluster centers. The frequency of these outliers among the full holdout set were compared against the expected frequency of 5% to determine statistical significance (chi-square, α=0.05). This test was performed for the in-time holdout, out-of-time holdout, and the combined holdout sets.

	In-time holdout	**Out-of-time holdout**	**Combined holdout**
1	6 / 105 (5.7%)	2 / 115 (1.7%)	8 / 220 (3.6%)
2	10 / 201 (5.0%)	15 / 240 (6.2%)	25 / 441 (5.7%)
3	3 / 42 (7.1%)	4 / 56 (7.1%)	7 / 98 (7.1%)
4	3 / 29 (10.3%)	2 / 57 (3.5%)	5 / 86 (5.8%)
5	0 / 20 (0.0%)	1 / 24 (4.2%)	1 / 44 (2.3%)
6	9 / 141 (6.4%)	11 / 193 (5.7%)	20 / 334 (6.0%)
7	1 / 28 (3.6%)	1 / 19 (5.3%)	2 / 47 (4.3%)
8	6 / 106 (5.7%)	1 / 79 (1.3%)	7 / 185 (3.8%)
9	3 / 30 (10.0%)	3 / 16 (18.8%)	6 / 46 (13.0%)
10	6 / 36 (16.7%)	1 / 38 (2.6%)	7 / 74 (9.5%)
11	2 / 47 (4.3%)	5 / 121 (4.1%)	7 / 168 (4.2%)
12	8 / 83 (9.6%)	1 / 23 (4.3%)	9 / 106 (8.5%)
13	2 / 91 (2.2%)	12 / 66 (18.2%)	14 / 157 (8.9%)
14	3 / 57 (5.3%)	0 / 23 (0.0%)	3 / 80 (3.8%)
15	0 / 2 (0.0%)	0 / 4 (0.0%)	0 / 6 (0.0%)
16	1 / 5 (20.0%)	2 / 17 (11.8%)	3 / 22 (13.6%)
17	3 / 43 (7.0%)	4 / 24 (16.7%)	7 / 67 (10.4%)
18	4 / 34 (11.8%)	0 / 16 (0.0%)	4 / 50 (8.0%)
19	0 / 18 (0.0%)	0 / 0 (NA)	0 / 18 (0.0%)
20	2 / 13 (15.4%)	0 / 0 (NA)	2 / 13 (15.4%)
Full	72 / 1131 (6.4%)	65 / 1131 (5.7%)	137 / 2262 (6.1%)
P-value	0.204	0.515	0.135

COVID-19 subgroup temporal analysis

The COVID-19 subgroup temporal analyses were performed only among patients with PDDs including COVID-19, sepsis, or viral pneumonia (N=5843). **Figure 4** is a chord diagram that depicts how patients transitioned between COVID-19 subgroups on a day-to-day basis throughout the course of hospitalization. Each link in the chord diagram represents patients transitioning from the source subgroup on one day (indicated at the beginning of each link) to the target subgroup on the following day (indicated at the end of each link). The link width is proportional to the number of observed transitions, e.g., the most common transitions are represented by the widest links. The link outline color represents the median number of days patients remained in the source subgroup before transitioning to the target subgroup, e.g., green or red outlines indicate patients spent 1 or 10+ day(s), respectively, in the source subgroup before transitioning to the target subgroup. To reduce complexity of the figure, we removed Subgroups 1, 3, 5, 15, and 16, as these accounted for fewer than 1% of transitions among these patients. **Table S9** quantitatively shows the transition counts and median durations, including all subgroups. All transitions with counts fewer than five were excluded from both Figure 4 and Table S9. Here, we describe a few transitions. The most common starting states (widest links originating from admission) were in Subgroups 6 and 2. Many of the patients in SG2 transitioned to either SG6 or SG8 after a median of 1 day. Most of the patients in SG6 were discharged after a median of 3 days or transitioned to SG12 after a median of 3 days. The most common subgroups that progressed to death or hospice were Subgroups 13, 17, and 18, with median durations of 5, 5, and 16 days, respectively.

Table 5 shows the top five complete paths throughout the hospitalization with 1) the highest frequency, 2) longest duration, 3) lowest mortality, and 4) highest mortality. There were 778 unique paths. 85 paths had at least 10 patients, accounting for 76.8% (4490/5843) of patients (**Table S10**). For each path, these tables show the constituent states with their median durations, total duration for the entire path, and mortality (including death and discharge to hospice) among all patients with the same base path (the path leading up to discharge or death). For the longest duration, only paths with at least 10 patients are shown. For the lowest and highest mortality rates, only paths with at least 50 and 10 patients, respectively, with the base path are shown.

Table 5. Top paths of interest. The top five paths with the highest frequency, longest duration, lowest mortality, and highest mortality among patients with a primary discharge diagnosis related to COVID-19 (5843 patients). The median number of days spent in each state before transitioning to the following state is shown in parentheses. Total Duration: median and interquartile range of hospital duration. Count: number (percentage) of patients characterized by the path. Mortality: percentage of patients who died or were discharged to hospice.

State 1	State 2	State 3	State 4	State 5	Total Duration (days)	Count (%)	Mortality (%)
Most Frequent							
6 (4.0 days)	Discharged				4.0 [3.0, 5.0]	856 (14.7%)	30 / 1030 (2.9%)
6 (3.0 days)	12 (4.0 days)	Discharged			8.0 [6.0, 11.0]	344 (5.9%)	37 / 498 (7.4%)
2 (1.0 days)	6 (2.0 days)	Discharged			4.0 [3.0, 5.0]	214 (3.7%)	4 / 276 (1.4%)
8 (4.0 days)	Discharged				4.0 [3.0, 6.0]	189 (3.2%)	21 / 325 (6.5%)
6 (1.0 days)	8 (3.0 days)	Discharged			4.0 [3.0, 7.0]	157 (2.7%)	16 / 254 (6.3%)
Longest Duration							
6 (3.0 days)	12 (4.0 days)	17 (15.0 days)	18 (44.0 days)	Care	64.0 [48.0, 95.0]	13 (0.2%)	10 / 25 (40.0%)
13 (1.0 days)	17 (11.0 days)	20 (53.0 days)	Care		62.0 [49.0, 77.0]	13 (0.2%)	2 / 15 (13.3%)
13 (2.0 days)	17 (15.5 days)	18 (43.5 days)	Care		61.0 [46.25, 76.0]	20 (0.3%)	10 / 30 (33.3%)
8 (1.5 days)	17 (17.0 days)	18 (34.5 days)	Care		57.0 [40.25, 69.0]	10 (0.2%)	10 / 20 (50.0%)
13 (2.0 days)	17 (9.0 days)	19 (43.5 days)	Care		55.0 [45.75, 80.25]	14 (0.2%)	12 / 26 (46.2%)
Lowest Mortality							
2 (1.0 days)	6 (3.0 days)				4.0 [3.0, 6.0]	276 (4.7%)	4 / 276 (1.4%)
6 (4.0 days)					4.0 [3.0, 5.0]	1030 (17.6%)	30 / 1030 (2.9%)
6 (3.0 days)	11 (5.0 days)				9.0 [6.0, 13.0]	157 (2.7%)	9 / 157 (5.7%)
6 (1.0 days)	8 (3.0 days)				5.0 [3.0, 8.0]	254 (4.3%)	16 / 254 (6.3%)
8 (5.0 days)					5.0 [3.0, 7.0]	325 (5.6%)	21 / 325 (6.5%)
Highest Mortality							
8 (4.0 days)	14 (4.0 days)	17 (4.0 days)			17.0 [12.0, 20.0]	11 (0.2%)	11 / 11 (100.0%)
6 (1.0 days)	8 (4.0 days)	17 (9.0 days)			17.0 [9.5, 22.0]	11 (0.2%)	9 / 11 (81.8%)
8 (5.0 days)	14 (3.0 days)	17 (11.0 days)	18 (31.0 days)		53.0 [43.0, 113.0]	13 (0.2%)	10 / 13 (76.9%)
8 (3.0 days)	17 (7.0 days)				9.5 [5.0, 14.0]	34 (0.6%)	26 / 34 (76.5%)
2 (1.0 days)	13 (1.0 days)	17 (5.0 days)			10.0 [6.0, 16.0]	15 (0.3%)	11 / 15 (73.3%)

In the most common path, 14.7% (856/5843) of patients spent a median of 4 days in SG6 before being discharged. In the path with the longest total duration, 0.2% (13/5843) of patients spent a median of 3 days in SG6, 4 days in SG12, 15 days in SG17, and 44 days in SG18 before being discharged to other services for additional care; median 64 days total. In the path with the lowest mortality (4/276, 1.4%), patients spent 1 day in SG2, 3 days in SG6, and were discharged. In the path with the highest mortality (11/11, 100%), patients spent a median of 4 days in SG8, 4 days in SG14, and 4 days in SG17 before passing away or being discharged to hospice.

Table S11 shows the mortality rate (death or discharge to hospice) among patients whose first day of hospitalization was classified in each subgroup; subgroup mortality was compared against the subgroup’s complement (chi-square; α=0.0025). The full cohort experienced 11.1% mortality. Patients starting in Subgroups 1-4 had significantly lower mortality (0%, 7.1%, 0.7%, and 2.1%, respectively), while patients starting in Subgroups 8, 10, 11, and 13 had significantly higher rates (19.2%, 66.7%, 40.9%, and 40.1%, respectively).

DISCUSSION

COVID-19 clinical subgroup analysis

Figure 3 shows that K-means clustering achieved good separation of patient vectors, as visualized by t-SNE. Some subgroups, especially those with lower severity (e.g., Subgroups 2, 4, 6, 8), appeared diffuse with patient vectors spread out across the plot. Other subgroups, especially those with higher severity (e.g., Subgroups 16-20), were well localized. Still, the t-SNE plot shows a few clusters of patients that may be visually separated from other nearby patients, but which were not isolated by K-means with K=20 (e.g., the far-left clusters within SG3). We visually evaluated the t-SNE plots with larger K but felt that the additional separation did not warrant the added complexity and decreased power to detect differences in the subgroups. Subgroup severity tended to increase toward the right of the t-SNE plot.

COVID-19 clinical subgroup characteristics

This section provides a condense description of the COVID-19 clinical subgroups using summary data from Table 1, COVID-characterization window prevalence rates from Tables 2 and S6, baseline-characterization window prevalence rates from Tables 3 and S7, and PDDs from Table S8. For brevity, throughout this section, we do not call out these tables when data are referenced. A more comprehensive description is provided in the Supplementary Materials.

SG1 had 740 (8.2%) patients. These patients were younger (median [IQR]: 31.3 [26.9-35.8] years) and 98.8% female. At baseline, SG1 had many prior visits (15 [9-22] visits) and 88.6% prevalence in pregnancy concepts. Within the COVID-characterization window, COVID-related concepts had low prevalence, and severity was 98.2% mild-moderate. Their hospital PDDs were mostly related to pregnancy. Thus, SG1 primarily represented patients with pregnancy-related hospitalizations and non-severe COVID-19.

SG2 was the largest, with 1530 (16.9%) younger (56.7 [34.9-72.0]) patients. At baseline, SG2 had lower prevalence of type-2 diabetes mellitus (T2DM), chronic kidney disease (CKD), and dementia. Within the COVID-window, COVID-related concepts had lower prevalence, severity was 95.3% mild-moderate, and PDDs were mostly not COVID-related. The patient vectors appeared dispersed in the t-SNE, signaling heterogeneity within the group. Thus, SG2 likely represented patients primarily hospitalized for a variety of other non-COVID conditions who had non-severe SARS-CoV-2 infections.

SG3 had 372 (4.1%) patients. Race-ethnicity had highest representation among White patients (57.0%). Vaccination rate (8.9%) was the highest. At baseline, SG3 had more prior visits (7 [3-17] visits) and higher prevalence of neck and spine conditions, but lower prevalence of COVID-related concepts (e.g., T2DM, heart disease). Within the COVID-window, most COVID-related concepts had lower prevalence rates, and severity was 92.7% mild-moderate. 48.5% of PDDs were related to findings of the spinal region. The patient vectors appeared dispersed in the t-SNE plot. Thus, SG3 likely represented another group of patients primarily hospitalized for non-COVID conditions who had non-severe SARS-CoV-2 infections.

SG4 had 274 (3.0%) patients. These patients were children (6.9 [1.9-14.1] years). Within the COVID-window, most of the common COVID conditions were less prevalent, however, fever was more prevalent (60.5%), which may be enriched since parents often seek medical attention when young children have fevers. Severity was mixed, with 66.4% mild-moderate, 19.7% severe, and 13.9% critical. Only one third of their PDDs were related to COVID-19. Thus, SG4 represented young children who primarily had non-severe SARS-CoV-2 infections.

SG5 had 176 (1.9%) patients; 65.9% male. At baseline, they had higher rates of heart disease (55.7%) and atherosclerosis of coronary artery (28.4%). Within the COVID-window, common COVID conditions were less prevalent, but SG5 had the second highest rates of pleural effusion (71.0%) and atelectasis (56.2%), likely driven by the high rates of heart failure (35.2%) and baseline heart disease. Correspondingly, 86.1% of their PDDs were CVDs. Although the severity algorithm detected 25.0% severe and 18.2% critical cases, the observed hypoxemia, AHRF, ARDS, and ventilation codes detected by the phenotyping algorithm may have been attributable to cardiovascular surgery,[24,25] overestimating frequency of severe and critical cases. Thus, SG5 represented patients hospitalized for CVDs who likely had non-severe SARS-CoV-2 infections.

SG6 had 1225 (13.5%) patients. At baseline, this group had fewer prior visits (1 [0-8] visits) and lower prevalence of most risk factors, like T2DM (5.4%), CKD (4.2%), and heart disease (10.8%). Within the COVID-window, SG6 had higher than average prevalence of dyspnea (68.5%), cough (52.5%), fever (49.1%), and hypoxemia (47.2%), but lower rates of many high-severity conditions, including ARDS (3.1%), sepsis (5.1%), and death (1.8%). The severity distribution was 40.4% mild-moderate, 56.7% severe, and 2.9% critical. Hospitalizations were shorter (4.0 [3.0-6.0] days) and PDDs were primarily related to COVID-19. Thus, SG6 primarily represented patients who had fewer risk factors and mostly developed mild-moderate or severe cases of COVID, but few critical cases and low mortality.

SG7 had 242 (2.7%) patients; 63.2% male. At baseline, SG7 had the most prior visits (58.0 [34.2-86.0]) and the most enriched set of baseline conditions, including solid organ transplant (SOT, 86.4%), immunodeficiency disorder (29.8%), T2DM (53.3%), CKD (67.8%), etc. Within the COVID-window, SG7 had higher rates of diarrhea, pleural effusion, and ESRD relative to other COVID patients, but these were similar to the patients' baseline prevalence; only ARFS was higher than at baseline (63.6% vs. 40.5%). The top PDDs were COVID-19 (35.5%) and complication of transplanted lung (12.8%). Despite the history of transplants and immunocompromised status, half of SG7 had mild-moderate cases, while the other half were severe or critical. A review article on COVID-19 in immunocompromised hosts found that patients with SOTs were predominantly older, male patients with comorbidities including hypertension, diabetes, CVD, CKD, and obesity (corresponding with our results), and symptoms including fever, dry cough, and diarrhea.[26] Fever and cough were present in SG7 but at similar rates to other COVID-19 inpatients. SG7 had the highest rate of diarrhea, but diarrhea was elevated at baseline. The review found risks for mechanical ventilation and mortality ranged from 5-67%, whereas SG7 had similar rates of ventilation (10.3%) and death (6.6%) to other patients.

SG8 had 831 (9.2%) older (68.3 [57.9-77.9]) patients. At baseline, they had elevated rates of T2DM (47.4%), essential hypertension (43.3%), and hyperlipidemia (31.5%). Within the COVID-window, SG8 had higher rates of dyspnea (56.4%), cough (37.3%), and asthenia (23.3%), but lower rates of ARDS (5.2%), sepsis (5.5%), and ESRD (0.5%). SG8 was also recorded with 70.6% essential hypertension and 65.7% T2DM, higher than observed within the baseline-window, likely because more than 25% of these patients had no visits within the baseline-window. The severity distribution was 43.2% mild-moderate, 49.6% severe, and 7.2% critical. Hospitalizations were shorter (5 [4-7] days), and their PDDs were primarily COVID-related. SG8 had similar prevalence rates of COVID-window conditions to SG6, but SG8 had higher prevalence of ARFS (33.5% vs. 15.8%), likely due to risk factors, including T2DM, age, and hypertension.[5] Thus, SG8 appeared to represent patients with T2DM who experienced mostly non-severe and severe cases of COVID-19.

SG9 had 263 (2.9%) older (64.6 [56.0-75.2]), 66.5% male patients. These patients had the highest representation among Black (31.2%) and lowest among White (17.1%) patients. At baseline, SG9 had more prior visits (15 [2-39] visits) and high rates of renal and cardiovascular diseases, including CKD (73.4%), ESRD (59.3%), heart disease (61.6%), atherosclerosis of coronary artery (44.1%), and T2DM (57.8%). Within the COVID-window, ESRD (80.6%) and heart failure (45.2%) were more prevalent than in other patients and elevated relative to baseline. Hospitalizations were longer (9 [6-14] days), but only half their PDDs were COVID-related. Thus, SG9 primarily represented patients with a history of kidney disease, CVDs, and T2DM whose severity distribution was similar to other patients, but with exacerbations to their underlying conditions (ESRD and heart failure).

SG10 had 263 (2.9%) patients. At baseline, SG10 had elevated rates of many conditions, including anemia, ARFS, cirrhosis, etc. Within the COVID-window, SG10 had higher rates of many conditions relative to the other patients, but these were also elevated at baseline: ARFS (COVID-window: 47.1%; baseline-window: 21.3%), tachycardia (35.4%; 19.4%), pleural effusion (35.0%; 13.7%), anemia (30.8%; 25.1%), atelectasis (29.3%; 13.7%), and sepsis (26.6%; 13.3%). Their hospitalizations were predominantly not COVID-related. SG10's patient vectors appeared dispersed (Fig. 3). Thus, SG10 appeared to be composed of a heterogeneous group of patients with a variety of underlying conditions which may have been exacerbated by SARS-CoV-2 infection.

SG11 had 349 (3.9%) older (71.8 [59.0-84.2]) patients. Within the COVID-window, SG11 had higher rates of dyspnea (63.0%), hypoxemia (64.2%), and AHRF (47.3%). Severity was 23.8% mild-moderate, 61.3% severe, and 14.9% critical. Hospitalizations were longer (9 [6-15] days), and PDDs were mostly COVID-related. SG11 represented older patients who mostly developed severe cases.

SG12 had 650 (7.2%) patients. At baseline, prevalence rates were lower for conditions, including T2DM (6.5%), heart disease (13.4%), and CKD (4.9%). Within the COVID-window, SG12 had higher rates of some conditions, like dyspnea (80.3%) and AHRF (67.8%), but lower rates of others, like sepsis (22.1%), pleural effusion (6.6%), ESRD (0.3%), etc. Their severities were mostly severe (70.3%). Hospitalizations occurred mostly within the first wave, were longer (9 [6-13] days), and PDDs were predominantly related to COVID-19. SG12 primarily represented patients from the first wave with lower prevalence of baseline conditions who developed severe cases of COVID-19.

SG13 had 720 (8.0%) patients. This was the oldest subgroup (82.7 [74.0-89.4]). At baseline, SG13 had higher prevalence of essential hypertension (43.9%), hyperlipidemia (31.4%), heart disease (30.8%), and dementia (25.3%). Within the COVID-window, SG13 had higher ARFS (52.9%), AHRF (45.8%), and death (33.3%), but lower invasive ventilation (4.0%). Along with 42.6% dementia within the COVID-window, these patients experienced altered mental status (33.8%) and delirium (19.7%). The severity was split: 34.7% mild-moderate, 27.4% severe, and 37.9% critical. Hospitalizations mostly occurred in the first two waves, were longer (7 [4-10] days), and their PDDs were mostly COVID-related. Compared to subgroups with higher severity, SG13, with the oldest patient population, exhibited lower prevalence of many COVID-related conditions and invasive ventilation but higher mortality.

SG14 had 454 (5.0%) older (70.2 [60.8-78.3] years), 61.9% male patients. SG14 had the highest representation among Hispanic patients (42.5%). At baseline, SG14 had higher prevalence of T2DM and hyperlipidemia, and within the COVID-window, baseline conditions were recorded at higher rates: T2DM (78.0%), hyperlipidemia (55.9%), hypertension (70.5%), and CKD due to T2DM (28.9%). Additionally, they had high prevalence of dyspnea (76.2%), AHRF (74.2%), and ARFS (52.2%), but low ESRD (0.2%) and invasive ventilation (1.1%). Their severities were mostly severe and critical. Their hospitalizations occurred during the first two waves, lasted 12 [8-17] days, and were predominantly COVID-related. SG14 mostly represented older and more commonly male patients with T2DM, hypertension, and some CKD, who developed severe and critical COVID-19 with ARFS but not ESRD.

SG15 had 43 (0.5%) patients. These patients were children (9.7 [4.2-15.7] years) and 62.8% male. Within the COVID-window, we saw higher rates of fever (69.8%), ARDS (65.1%), pleural effusion (67.4%), atelectasis (76.7%), and invasive ventilation (65.1%). 44.2% had heart failure, the second highest rate, and 23.3% had congenital heart disease, likely predisposing them to heart failure. Their severities were mostly critical and severe. Hospitalizations were much longer (45.0 [22.5-81.0] days). Only 30.2% of PDDs were COVID-19, but 44.2% were cardiac disorders. SG15 represented children who primarily developed critical SARS-CoV-2 infections, many of whom had congenital heart disease and were primarily hospitalized for their CVDs, often leading to pleural effusion and atelectasis.

SG16 had 58 (0.6%) patients; 65.5% male. Within the COVID-window, SG16 had high AHRF (79.3%), ARDS (65.1%), pleural effusion (82.8%), atelectasis (50.0%), heart failure (36.2%), ARFS (81.0%), septic shock (56.9%), invasive ventilation (82.8%), and death (41.4%). Severity was 93.1% critical. Their hospitalizations lasted 41.0 [20.0-62.8] days and were mostly COVID-related. SG16 represented patients who developed critical COVID-19 with long hospitalizations and high rates of pleural effusion, atelectasis, heart failure, and mortality.

SG17 had 367 (4.1%) older (68.1 [57.9-76.8]) patients. Within the COVID-window, SG17 had elevated prevalence of AHRF (86.1%), ARDS (68.1%), pleural effusion (41.4%), heart failure (24.0%), ARFS (75.2%), and septic shock (50.1%). 80.4% had invasive ventilation and 64.6% passed away, the highest rate among the subgroups, despite age being the only significant risk factor at baseline and other groups exhibiting more critical complications. Their severity was critical (94.6%). Their hospitalizations occurred mostly during the first two waves, were longer (13.0 [7.0-19.5] days), and their PDDs were COVID-related. Similar to SG13, SG17 primarily represented older patients who, despite not having significantly higher prevalence rates of risk factors at baseline other than age, continued to develop severe or critical cases of COVID-19 and experienced very high mortality. While SG17 was younger than SG13, SG17 had higher prevalence rates of critical complications (e.g., ARDS 68.1% vs. 11.4%, ARFS 75.2% vs. 52.9%, septic shock 50.1% vs. 6.9%) and mortality (64.6% vs. 33.3%).

Subgroups 18, 19, and 20 generally have similar characteristics and are described together. Subgroups 18-20 had 220 (2.4%), 160 (1.8%), and 114 (1.3%) patients, respectively. They were predominantly male (60.9-68.8%). Within the COVID-window, Subgroups 18-20 commonly had elevated rates of high-severity concepts, including ARDS (87.5-100.0%), pleural effusion (43.0-77.3%), ARFS (85.6-89.5%), septic shock (70.2-79.5%), and altered mental status (39.4-46.5%). Severities were nearly all critical. SG19 had hospitalizations during the first two waves and SG20 during the first wave. Their PDDs were mostly COVID-related. In general, aside from having more male patients, these subgroups were not significantly different from other patients at baseline, but they developed critical cases of COVID-19 with long hospitalizations (40.0-46.0 days median), invasive ventilation (96.4-100%), and high mortality (26.3-45.0%). Despite the similarities between Subgroups 18-20, Fig. 3 shows the subgroups were well separated. Compared to the other two subgroups, SG20 had the lowest rates of heart failure, pleural effusion, and atelectasis, and although SG19 had the highest rates of heart failure and atelectasis, SG18 had the highest rate of pleural effusion. Given the different stages of the pandemic over which these subgroups occurred, there likely were also differences in the clinical care captured by the patient vectors, but which were not presented in this subgroup characterization.

Looking at vaccination rates across the subgroups, we see a trend with higher rates among the lower severity subgroups and lower vaccination rates among the higher severity groups. However, subgroup vaccination rates were also affected by when the hospitalizations within the subgroup occurred and the availability of the vaccine. For example, Subgroups 4, 15, and 20 had no patients vaccinated, but these groups primarily had their hospitalizations prior to FDA vaccine emergency use authorization for their constituents, e.g., children in Subgroups 4 and 15, and first-wave pandemic patients in SG20. Patient vaccination status was also likely underestimated since vaccination data were only collected from CUIMC EHR and NYC registry; thus, non-NYC residents who were vaccinated in other healthcare systems were not captured by these sources.

Clinical subgroup generalizability

The in-time and out-of-time holdout evaluations showed good generalizability of the identified COVID-19 clinical subgroups to patients whose data were not seen during training of the K-means clustering algorithm (Table 4). As expected, there was some variance in the frequency of outliers within the small sample sizes of each subgroup, but among each of the holdout sets, the outlier frequency was not significantly different from the expected frequency of 5.0%. The out-of-time holdout evaluation showed that clinical subgroups were representative of the most recent hospitalizations (July 20, 2021 – December 1, 2021). However, Subgroups 19 and 20, which had hospitalizations primarily in the first wave of the pandemic, were not assigned to any patients in the out-of-time holdout set, suggesting that changes in clinical practice and/or recording may have made these subgroups obsolete. Despite this, we observed similar outlier frequencies between the out-of-time and in-time holdout sets (5.7% vs. 6.4%, P=0.600), suggesting that the subgroups could represent new patients as well as those contemporaneous to the patients in the training data. Generalizability has not yet been tested against patient data from other institutions or geographic locations.

COVID-19 subgroup temporal analysis

Subgroup transition analysis

The chord diagram in Fig. 4 depicts day-to-day transitions between COVID-19 subgroups when the patient vectors were generated per day of hospitalization among patients with PDDs related to COVID-19. This format helps visualize how the COVID-19 clinical subgroups were interrelated and facilitates recognition of the most common transitions and their durations, while Table S9 provides the quantitative details. For example, the most common starting state was SG6, which primarily represented patients who had fewer risk factors and primarily developed mild-moderate (40.4%) or severe (56.7%) cases of COVID-19. Thus, these patients likely already had moderate or severe COVID-19 when they were admitted to the hospital. The most common transition from SG6 was for these patients to be discharged after a median of 3 days, corresponding with the complete path (SG6→discharge) being the most common path (Table 5). Patients who were not discharged from SG6 commonly transitioned to SG12 after a median of 3 days. SG12 had more severe cases of COVID-19 with higher prevalence rates of AHRF (67.8% vs. 35.3%), dyspnea (80.3% vs. 68.5%), and ARFS (28.6% vs. 15.8%), suggesting that these patients' conditions worsened approximately three days after admission. Fig. 4 also shows a general trend that the lower subgroups (e.g., Subgroups 2-13) had relatively quick transitions into them (1-3 days), as seen by the green-yellow outlines of the inbound links. In contrast, patients in the higher subgroups (e.g., Subgroups 14-20) tended to remain in these states for long durations (7+ days, orange-red outlines) before transitioning out.

Few patients transitioned into the most severe subgroups (Subgroups 18-20) without first passing through SG17. SG17 had much shorter hospitalizations (median 13 days vs. 40-46 days) and lower rates of ARDS (68.1% vs. 87.5-100%), sepsis (56.9% vs. 80.0-81.9%), septic shock (50.1% vs. 70.2-79.5%), invasive ventilation (80.4% vs. 96.4-100%), etc. However, SG17 had the highest mortality (64.6% vs. 26.3-45.0%). This may suggest the risk of death peaked when patients were in SG17, and patients who survived this high-risk period may enter a state with more complications and longer hospitalizations but lower risk of death.

Subgroup path analysis

Table S10 shows the complete paths throughout the hospitalization, and Table 5 shows a sample of interesting paths with 1) the highest frequency, 2) longest duration, 3) lowest mortality, and 4) highest mortality. In the most common path, 14.7% (856/5843) of patients spent a median of 4 days in SG6 before being discharged, as described above.

The five paths with the longest total durations shared some commonalities. In all five paths, patients spent 9.0-17.0 days (median durations) in SG17 as their penultimate state before transitioning into one of Subgroups 18-20 as the last state before being discharged. This last state was where these patients spent for the majority of their hospitalization, with median durations ranging from 34.5-53.0 days. Even after their lengthy hospitalizations, these patients were discharged to additional care services, indicating that they still required professional care.

In the path with the lowest mortality, patients spent a median of 1 day in SG2 and 3 days in SG6 before being discharged. The mortality and hospice rate was only 1.4% (4/276) among patients who experienced this base path. Surprisingly, the path with only a single state in SG2 had higher mortality (13/171, 7.6%), and death occurred sooner (median 2 days) despite SG2 having slightly lower mortality than SG6 (1.3% vs. 1.8%, Table 2) in the clinical subgroup analysis. This may be because the clinical subgroup analysis included all COVID-19 hospitalizations, whereas this subgroup temporal analysis only included patients with COVID-19 PDDs. The PDD criterion removed 93.9% of the patients from SG2, which may have substantially shifted the composition of SG2.

In the five base paths with the highest mortality, mortality ranged from 73.3% (11/15) to 100% (11/11) among the paths, though each path had small sample sizes. Four of these paths ended on SG17 and had medium hospital lengths of stay (median durations ranged from 9.5-17.0 days). In the fifth path, patients passed through SG17 before transitioning to SG18 and ultimately passing away after an extended hospitalization (median 53.0 days). From the clinical subgroup analysis, SG17 and SG18 were also observed with the highest mortality, 64.6% and 45.0%, respectively (Table 2). With four of the highest mortality paths ending on SG17, this path analysis corroborates the previous observation that the risk of death may have peaked when patients were in SG17.

Admission state mortality analysis

Table S11 shows the mortality rates among patients whose first day of hospitalization were classified into each of the subgroups. Patients starting in Subgroups 1-4 had significantly lower mortality rates, while patients starting in Subgroups 8, 10, 11, and 13 had significantly higher mortality rates. Although the COVID-19 subgroups were defined by applying a clustering algorithm on patient vectors derived from clinical data covering the entire COVID-characterization window, the patient vectors and corresponding subgroup predictions based on data from the admission day captured enough information to differentiate patients into subgroups associated with varying risk for mortality. This suggests that the patient vectors containing admission data alone may be useful for predicting prognosis.

Research in context

A few studies were previously published looking at subgroups of hospitalized COVID-19 patients. Lusczek et al. analyzed EHR data from the first 72 hours of hospitalization of 1,022 COVID-19 patients and found three clinical phenotypes associated with various comorbidities and outcomes. Phenotype-I patients (23.1%) were older and commonly had renal, hematologic, and cardiac comorbidities. Phenotype-II patients (60%) were associated with moderate severity. Phenotype-III patients (16.9%) were more often female and had respiratory comorbidities. Compared to Phenotype-III, Phenotypes I and II had increased odds of complications (respiratory, renal, and metabolic) and worse outcomes (ICU admission, ventilation, hospital LOS, and death). Oh et al. analyzed temporal patterns in EHR data from the first 24 hours of ICU admission for 1036 patients and identified four clinical subphenotypes. Subphenotype I patients (22.5%) had rapid respirations and heartbeat and enjoyed a relatively good prognosis with less need for invasive interventions. Subphenotype II patients (40.3%) were younger and had the fewest abnormal biomarker levels, low mortality, and the highest probability of being discharged. Subphenotype III patients (25.0%) were older, more likely male, and experienced clinical deterioration during the first 24 hours of ICU admission, leading to poor outcomes. Subphenotype IV patients (12.2%) nearly all experience ARDS and required mechanical ventilation within 24 hours of ICU admission.

This study performed a deeper dive on more precisely defined subgroups by analyzing EHR data from a larger data collection window (minimally including the entire hospitalization) on 9051 patients and provided a highly detailed characterization of each subgroup’s demographics, baseline health, vaccination status, COVID-19 severity, conditions, outcomes, and visit details. Patients in several subgroups were hospitalized for underlying conditions (pregnancy, CVD, etc.) and only experienced mild-moderate COVID-19. SG7 included transplant recipients who mostly developed mild-moderate or severe disease. SG9 had a history of T2DM, kidney disease, and CVD, and suffered the highest rates of heart failure and ESRD. SG13 was the oldest (median: 82.7 years) and had mixed severity but high mortality (33.3%). SG15 included children, many of whom had congenital heart diseases and experienced critical cases of COVID-19 with high rates of pleural effusion and atelectasis. SG17 had critical disease and the highest mortality (64.6%), with age being the only notable risk factor. Subgroups 18-20 had critical disease with high rates of ARDS, ARFS, septic shock, and long LOS (median: 40+ days). We further analyzed how patients transitioned through these subgroups throughout the course of hospitalization to characterize temporal patterns between subgroups, including the duration (days) patients spent in each subgroup, common transitions between them, and subgroup paths with the highest prevalence, longest duration, and lowest and highest mortality.

The findings from this study could be used in many ways to support clinical decision making. Understanding these subgroups may help clinicians triage patients for better management and earlier intervention. The evaluation of the assigned COVID-19 subgroup at hospital admission suggests that the identified subgroups may be predictive of outcomes. Therefore, the patient embeddings have potential for being used to predict risks of patients developing complications like ARFS or atelectasis, which can further inform the level of vigilance required for monitoring high-risk patients, tailoring treatment plans to reduce these risks, and coordinating care across specialties when patients are at risk of multiple organ failure. Data from the visit characteristics could be leveraged for planning hospital resource utilization, especially for subgroups with long hospitalizations or for those who ultimately transfer for skilled nursing facilities and other care services. Clinical trials could possibly stratify participants according to these subgroups to determine whether the intervention is more effective or has fewer adverse outcomes in certain subgroups.

Limitations

There are several limitations to this study. The K-means clustering method required manual selection of the number of clusters; thus, the number of COVID-19 subgroups identified was influenced by subjective decision. Prior to selecting K-means clustering for this study, we evaluated several clustering methods on a preliminary data set, some of which allow clustering to be performed without manual selection of hyperparameters. With the other tested algorithms, we found either patients poorly distributed over the resulting clusters (e.g., most patients in a single cluster or many clusters with a single patient), or clusters poorly separated on t-SNE, thus, we ultimately selected K-means for this study. The elbow method to guide the selection of K yielded a broad curve which only helped eliminate lower and upper extreme values from consideration. We selected K=20 upon visual inspection of the t-SNE plot, prior to performing subgroup characterization. However, the subgroup characterization revealed that we may have been slightly too conservative in our choice, as several groups appeared to contain heterogeneous sets of subpopulations, such as Subgroups 2-4, which may have benefitted from a larger selection of K to differentiate their subpopulations. Additionally, the use of Euclidean distance in K-means clustering is not ideal for the high-dimensional patient vectors. Although we found the clusters identified by K-means to be meaningful, other clustering methods may produce better results.

The inclusion criteria for this study could have been adjusted to make the results more specific to COVID-19. As seen in Table S8, many of the included patients did not have COVID-19 as their primary discharge diagnosis, indicating that these patients were primarily hospitalized for other health reasons and may have been asymptomatic or mildly symptomatic of COVID-19. However, by not requiring the COVID-19 PDD in the subgroup analysis, the resulting COVID-19 subgroups are more representative of all patients who test positive while hospitalized, including those with non-severe cases, which may have been excluded if the COVID-19 PDD was required. For prospective application, these classifications need to be applicable to all SARS-CoV-2 positive patients since the PDDs will not be known until the patient is discharged.

This study was performed using observational EHR data. EHR data are collected as part of routine clinical care, as opposed to being collected systematically for research purposes, resulting in EHR data commonly having known data quality issues and biases.[27–29] For example, individually, race and ethnicity were not well captured in our EHR data; we combined race and ethnicity into a single variable to minimize the impact of missingness in the individual variables. Healthcare processes can influence EHR data in many ways, e.g., recorded conditions and procedures are biased by reimbursement incentives, laboratory measurements exhibit diurnal variations due to overnight measurements ordered on sicker patients, etc.[30] In several subgroups, we could see chronic conditions being reported at higher prevalence rates within the COVID-characterization window than in the baseline-characterization window. This was likely an underestimation of baseline prevalence rates as the conditions could have been under-coded since many patients had no or few visits within the baseline-characterization window. Also, the EHR data analyzed in this study included data from the beginning of the pandemic through to December 2021, a time range where much changed, including changing best treatment practices, emergence of new SARS-CoV-2 variants, availability of vaccines, and fluctuating levels of stress to the hospital system through the waves of the pandemic. These changes were not controlled for in the subgroup analysis, and some subgroups can even be characterized based on these external factors, e.g., SG20 had most of its hospitalizations within the weeks after the first confirmed case in New York.

Finally, the results of this study have not been externally validated yet. The National COVID Cohort Collaborative (N3C) has collected EHR data on over 12 million patients with over 4 million SARS-CoV-2-positive cases from over 50 data partners with released data. N3C data are stored in the same Observational Medical Outcomes Partnership (OMOP) CDM format as used in this study and present an opportunity for robust validation.

CONCLUSION

Using a patient embedding model and clustering methods, twenty subgroups of hospitalized COVID-19 patients were identified and labeled in order from lowest to highest severity. These subgroups varied widely according to their demographics, baseline health, complications, outcomes, severity, and temporal characteristics. This manuscript provides a highly detailed characterization of each subgroup and the common temporal transitions and paths between them throughout the course of hospitalization. The entire analysis pipeline, including transforming EHR data from OMOP into medical coding sequences and, subsequently, into patient vector representations and clinical characterization, is publicly available (see data availability statement). Future studies will investigate using these findings toward predictive analytics to improve outcomes for COVID-19 patients.

FUNDING

Research reported in this publication was supported by the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number OT2TR003434. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

AUTHOR CONTRIBUTIONS

Conceptualization, CNT, JEZ, and CW; Methodology, CNT, JEZ, PHC, YF, KN, and CW; Investigation, CNT, JEZ, and CW; Visualization, CNT; Funding acquisition, CNT and CW; Project administration, CNT and CW; Supervision, CW; Writing – Original Draft, CNT; Writing – Review & Editing, CNT, JEZ, PHC, YF, KN, and CW.

CONFLICT OF INTEREST STATEMENT

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

The data are not publicly available as the data were derived from electronic health records and include sensitive individual-level information that could compromise research participant privacy.

The code used in this analysis are available at: https://github.com/WengLab-InformaticsResearch/covid_subgrouping.

FIGURE LEGENDS

Fig. 1. Methodology overview. (A) The patient timeline depicts the COVID-19 cohort inclusion criteria and the definitions of the baseline-characterization and COVID-characterization windows. **(B)** Patient-level EHR data were extracted from an OMOP database and transformed into medical coding sequences (MCS). Unique concepts were recorded and shuffled for each day. **(C)** Patient MCSs were converted into vector embeddings using paragraph vector models (distributed memory and distributed bag-of-words; concatenated). **(D)** Patient vectors were clustered to identify COVID-19 clinical subgroups. **(E)** Subgroups were characterized by demographics, visit details, conditions and outcome prevalence rates, and temporal patterns between subgroups.

Fig. 2. COVID-19 cohort characteristics. Normalized distributions from the full COVID-19 inpatient cohort: **(A)** age, **(B)** race-ethnicity, **(C)** hospitalization length of stay, **(D)** hospitalization start date, **(E)** number of healthcare visits in the baseline-characterization window, and **(F)** number of healthcare visits in the COVID-characterization window. AIAN: American Indian and Alaska Native; NHPI: Native Hawaiian and Pacific Islander.

Fig. 3. **t-distributed stochastic neighbor embedding of COVID-19 patient vectors.** Each marker is a vector representation of a patient in the COVID-19 inpatient cohort; the symbol indicates the COVID-19 subgroup.

Fig. 4. COVID-19 subgroup transitions. The chord diagram depicts the transitions between COVID-19 subgroups among patients with primary discharge diagnoses related to COVID-19. A link between two groups indicates the transition from the source group to the target group. Link width is proportional to the number of transitions. Links are arranged clockwise in descending order of the transition count. Link outline color (colorbar legend) represents the median number

of days patients remained in the source subgroup before transitioning to the target subgroup. Link fill color indicates the group each transition originates from, matching the fill color of the arc representing each subgroup. For figure clarity, transitions with fewer than five counts were excluded, and Subgroups 1, 3, 5, 15, and 16 were excluded since few patients entered these groups. Care: patients discharged to additional care services; death: patients died or discharged to hospice.

REFERENCES

1 WHO COVID-19 Dashboard. World Health Organ. 2020.https://covid19.who.int (accessed 8 Apr 2022).

2 CDC. COVID Data Tracker. Cent. Dis. Control Prev. 2020.https://covid.cdc.gov/covid-data-tracker (accessed 8 Apr 2022).

3 Yuki K, Fujiogi M, Koutsogiannaki S. COVID-19 pathophysiology: A review. *Clin Immunol* 2020;**215**:108427. doi:10.1016/j.clim.2020.108427

4 Bader F, Manla Y, Atallah B, *et al.* Heart failure and COVID-19. *Heart Fail Rev* 2021;**26**:1–10. doi:10.1007/s10741-020-10008-2

5 Hirsch JS, Ng JH, Ross DW, *et al.* Acute kidney injury in patients hospitalized with COVID-19. *Kidney Int* 2020;**98**:209–18. doi:10.1016/j.kint.2020.05.006

6 Phipps MM, Barraza LH, LaSota ED, *et al.* Acute Liver Injury in COVID-19: Prevalence and Association with Clinical Outcomes in a Large U.S. Cohort. *Hepatology* 2020;**72**:807–17. doi:https://doi.org/10.1002/hep.31404

7 Nobel YR, Phipps M, Zucker J, *et al.* Gastrointestinal Symptoms and Coronavirus Disease 2019: A Case-Control Study From the United States. *Gastroenterology* 2020;**159**:373-375.e2. doi:10.1053/j.gastro.2020.04.017

8 Niazkar HR, Zibaee B, Nasimi A, *et al.* The neurological manifestations of COVID-19: a review article. *Neurol Sci* 2020;**41**:1667–71. doi:10.1007/s10072-020-04486-3

9 Zhou H, Lu S, Chen J, *et al.* The landscape of cognitive function in recovered COVID-19 patients. *J Psychiatr Res* 2020;**129**:98–102. doi:10.1016/j.jpsychires.2020.06.022

10 Argenziano MG, Bruce SL, Slater CL, *et al.* Characterization and clinical course of 1000 patients with coronavirus disease 2019 in New York: retrospective case series. *BMJ* 2020;**369**:m1996. doi:10.1136/bmj.m1996

11 Brat GA, Weber GM, Gehlenborg N, *et al.* International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. *Npj Digit Med* 2020;**3**:1–9. doi:10.1038/s41746-020-00308-0

12 Weber GM, Zhang HG, L'Yi S, *et al.* International Changes in COVID-19 Clinical Trajectories Across 315 Hospitals and 6 Countries: Retrospective Cohort Study. *J Med Internet Res* 2021;**23**:e31400. doi:10.2196/31400

13 Kostka K, Duarte-Salles T, Prats-Uribe A, *et al.* Unraveling COVID-19: A Large-Scale Characterization of 4.5 Million COVID-19 Cases Using CHARYBDIS. *Clin Epidemiol* 2022;**14**:369–84. doi:10.2147/CLEP.S323292

14 Morais IJ, Polveiro RC, Souza GM, *et al.* The global population of SARS-CoV-2 is composed of six major subtypes. *Sci Rep* 2020;**10**:18289. doi:10.1038/s41598-020-74050-8

15 Chen Z, Feng Q, Zhang T, *et al.* Identification of COVID-19 subtypes based on immunogenomic profiling. *Int Immunopharmacol* 2021;**96**:107615. doi:10.1016/j.intimp.2021.107615

16 Huang Y, Pinto MD, Borelli JL, *et al.* COVID Symptoms, Symptom Clusters, and Predictors for Becoming a Long-Hauler: Looking for Clarity in the Haze of the Pandemic. 2021. doi:10.1101/2021.03.03.21252086

17 Lusczek ER, Ingraham NE, Karam BS, *et al.* Characterizing COVID-19 clinical phenotypes and associated comorbidities and complication profiles. *PLOS ONE* 2021;**16**:e0248956. doi:10.1371/journal.pone.0248956

18 Sudre CH, Lee KA, Lochlainn MN, *et al.* Symptom clusters in COVID-19: A potential clinical prediction tool from the COVID Symptom Study app. *Sci Adv* 2021;**7**:eabd4177. doi:10.1126/sciadv.abd4177

19 Kenny G, McCann K, O'Brien C, *et al.* Identification of Distinct Long COVID Clinical Phenotypes Through Cluster Analysis of Self-Reported Symptoms. *Open Forum Infect Dis* 2022;**9**:ofac060. doi:10.1093/ofid/ofac060

20 Oh W, Jayaraman P, Sawant AS, *et al.* Using sequence clustering to identify clinically relevant subphenotypes in patients with COVID-19 admitted to the intensive care unit. *J Am Med Inform Assoc* 2022;**29**:489–99. doi:10.1093/jamia/ocab252

21 Le Q, Mikolov T. Distributed Representations of Sentences and Documents. In: *International Conference on Machine Learning*. 2014. 1188–96.http://proceedings.mlr.press/v32/le14.html (accessed 15 Apr 2019).

22 Klann JG, Estiri H, Weber GM, *et al.* Validation of an internationally derived patient severity phenotype to support COVID-19 analytics from electronic health record data. *J Am Med Inform Assoc JAMIA* 2021;**28**:1411–20. doi:10.1093/jamia/ocab018

23 Řehůřek R, Sojka P. Software Framework for Topic Modelling with Large Corpora. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: : ELRA 2010. 45–50.

24 Nakazato K, Takeda S, Tanaka K, *et al.* Aggressive treatment with noninvasive ventilation for mild acute hypoxemic respiratory failure after cardiovascular surgery: Retrospective observational study. *J Cardiothorac Surg* 2012;**7**:41. doi:10.1186/1749-8090-7-41

25 Rong LQ, Di Franco A, Gaudino M. Acute respiratory distress syndrome after cardiac surgery. *J Thorac Dis* 2016;**8**:E1177–86. doi:10.21037/jtd.2016.10.74

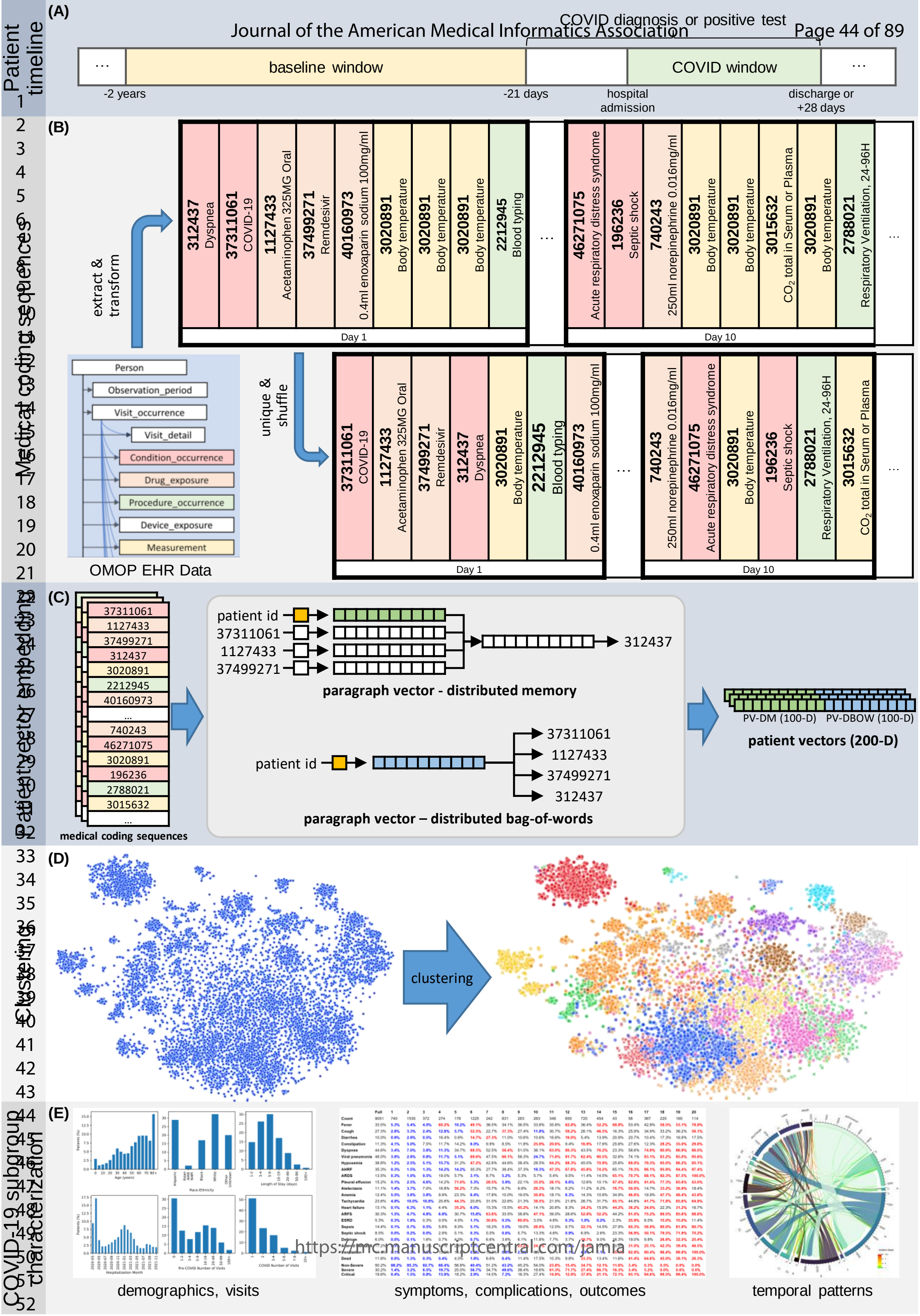
26 Fung M, Babik JM. COVID-19 in Immunocompromised Hosts: What We Know So Far. *Clin Infect Dis* 2021;**72**:340–50. doi:10.1093/cid/ciaa863

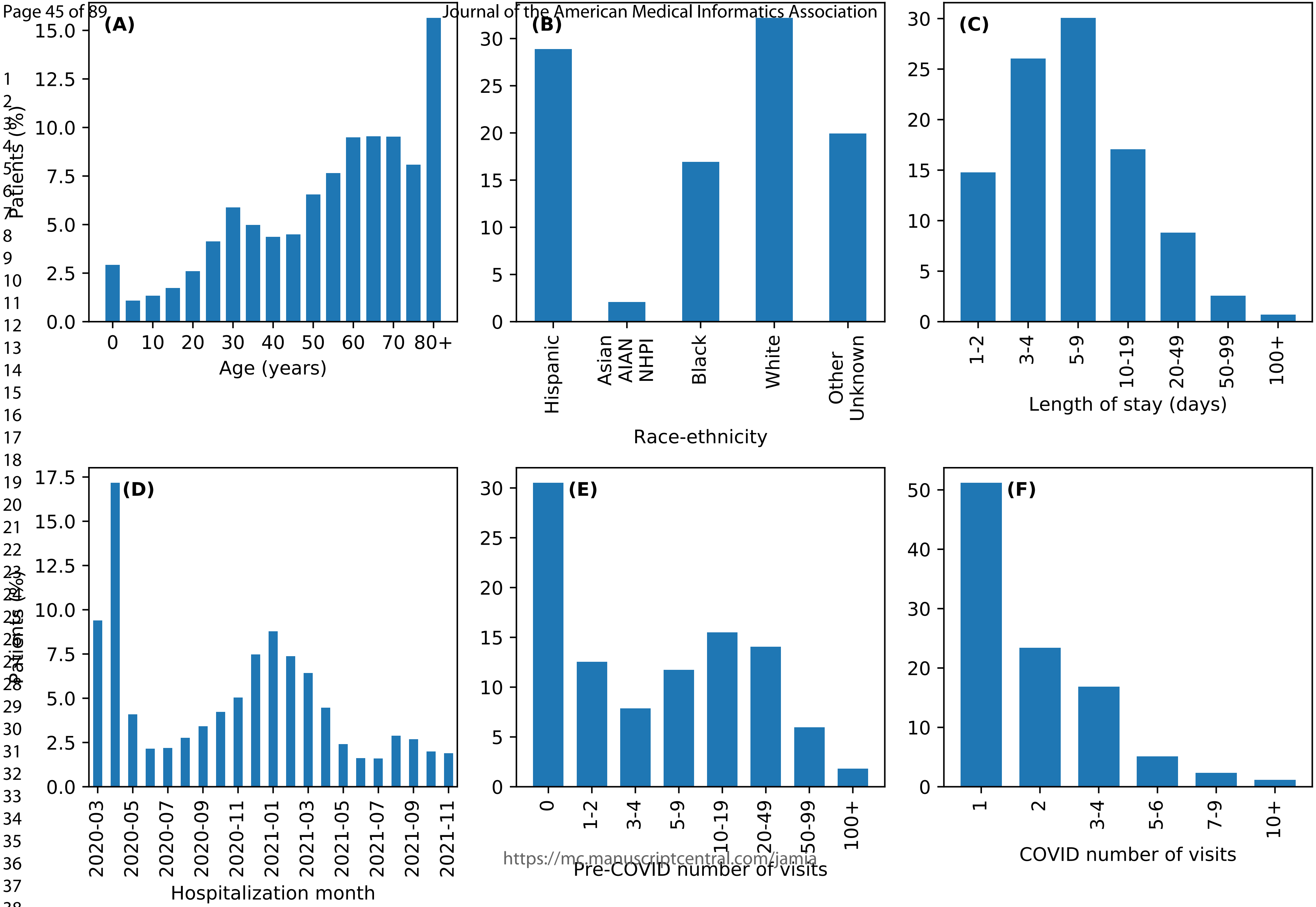
27 Ta CN, Weng C. Detecting Systemic Data Quality Issues in Electronic Health Records. *Stud Health Technol Inform* 2019;**264**:383–7. doi:10.3233/SHTI190248

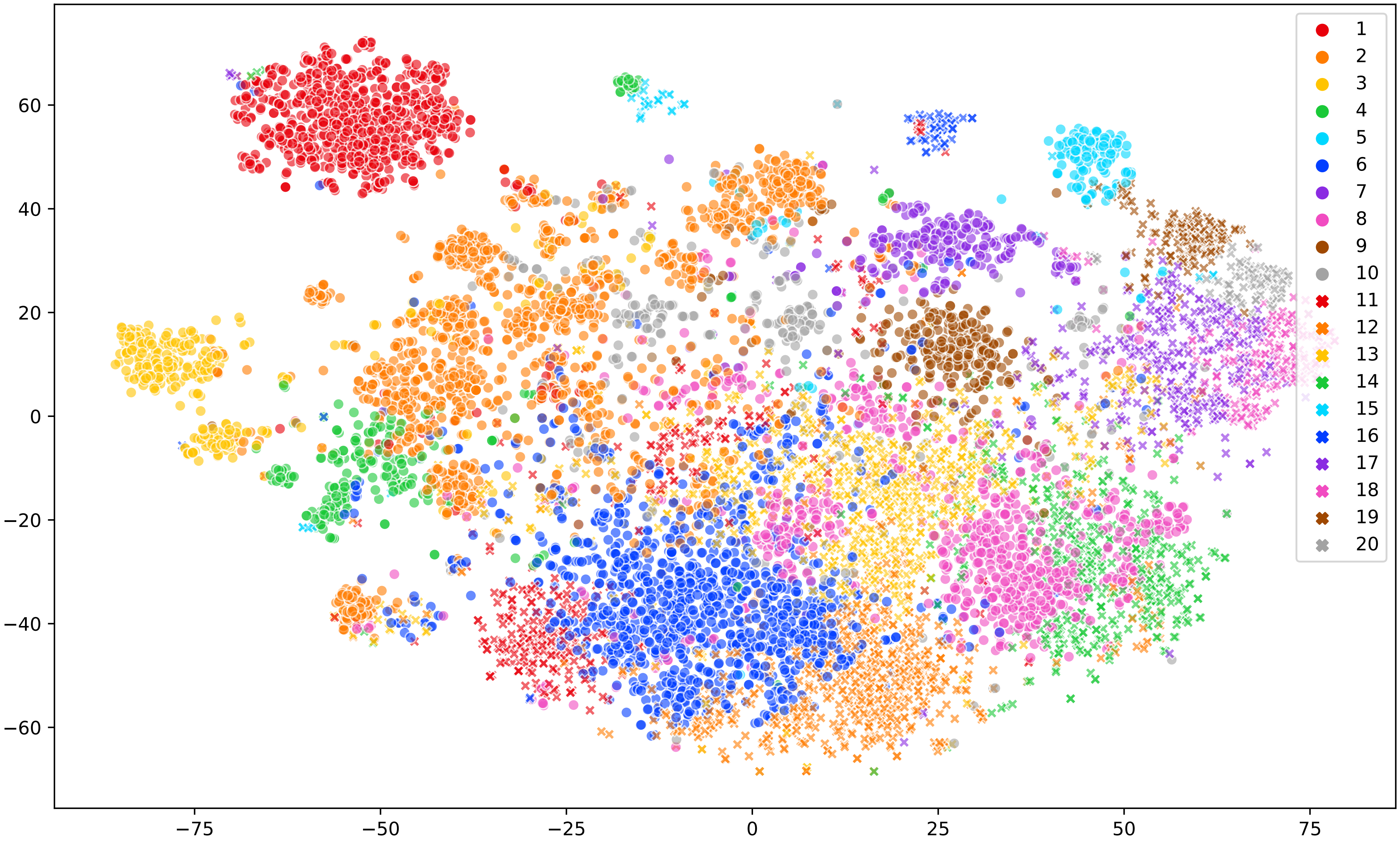
28 Weiskopf NG, Bakken S, Hripcsak G, *et al.* A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. *EGEMs Gener Evid Methods Improve Patient Outcomes* 2017;**5**:14. doi:10.5334/egems.218

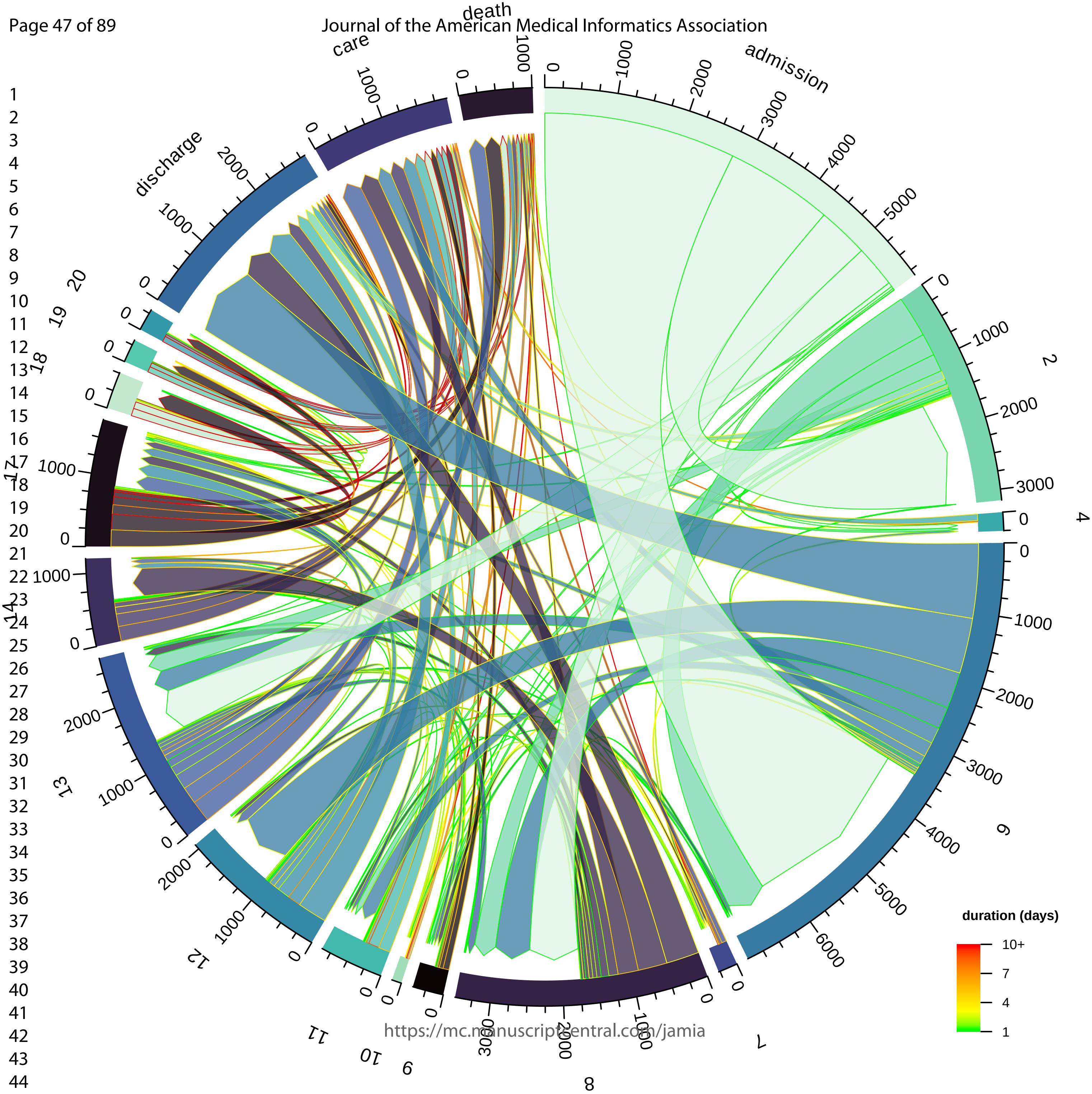
29 Kahn MG, Callahan TJ, Barnard J, *et al.* A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS Wash DC* 2016;**4**:1244. doi:10.13063/2327-9214.1244

30 Hripcsak G, Albers DJ. High-fidelity phenotyping: richness and freedom from bias. *J Am Med Inform Assoc JAMIA* Published Online First: 12 October 2017. doi:10.1093/jamia/ocx110


Patient timeline
(A)
COVID diagnosis or positive test
…
baseline window
COVID window
…
-2 years
-21 days
hospital admission
discharge or +28 days
Medical coding sequences
(B)
extract & transform
312437 Dyspnea
37311061 COVID-19
1127433 Acetaminophen 325MG Oral
37499271 Remdesivir
40160973 0.4ml enoxaparin sodium 100mg/ml
3020891 Body temperature
3020891 Body temperature
3020891 Body temperature
2212945 Blood typing
Day 1
…
46271075 Acute respiratory distress syndrome
196236 Septic shock
740243 250ml norepinephrine 0.016mg/ml
3020891 Body temperature
3020891 Body temperature
3015632 CO2 total in Serum or Plasma
3020891 Body temperature
2788021 Respiratory Ventilation, 24-96H
Day 10
…
Person
Observation_period
Visit_occurrence
Visit_detail
Condition_occurrence
Drug_exposure
Procedure_occurrence
Device_exposure
Measurement
OMOP EHR Data
unique & shuffle
37311061 COVID-19
1127433 Acetaminophen 325MG Oral
37499271 Remdesivir
312437 Dyspnea
3020891 Body temperature
2212945 Blood typing
40160973 0.4ml enoxaparin sodium 100mg/ml
Day 1
…
740243 250ml norepinephrine 0.016mg/ml
46271075 Acute respiratory distress syndrome
3020891 Body temperature
196236 Septic shock
2788021 Respiratory Ventilation, 24-96H
3015632 CO2 total in Serum or Plasma
Day 10
…
Patient vector embedding
(C)
37311061
1127433
37499271
312437
3020891
2212945
40160973
…
740243
46271075
3020891
196236
2788021
3015632
…
medical coding sequences
patient id
37311061
1127433
37499271
312437
paragraph vector - distributed memory
patient id
37311061
1127433
37499271
312437
paragraph vector – distributed bag-of-words
PV-DM (100-D)
PV-DBOW (100-D)
patient vectors (200-D)
Clustering
(D)
clustering
COVID-19 subgroup characterization
(E)
demographics, visits
symptoms, complications, outcomes
temporal patterns


(A)
Patients (%)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
0
10
20
30
40
50
60
70
80+
Age (years)
(B)
0
5
10
15
20
25
30
Hispanic
Asian
AIAN
NHPI
Black
White
Other
Unknown
Race-ethnicity
(C)
0
5
10
15
20
25
30
1-2
3-4
5-9
10-19
20-49
50-99
100+
Length of stay (days)
(D)
Patients (%)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07
2021-09
2021-11
Hospitalization month
(E)
0
5
10
15
20
25
30
0
1-2
3-4
5-9
10-19
20-49
50-99
100+
Pre-COVID number of visits
(F)
0
10
20
30
40
50
1
2
3-4
5-6
7-9
10+
COVID number of visits


60
40
20
0
−20
−40
−60
−75
−50
−25
0
25
50
75
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20


death
0
1000
care
0
1000
admission
0
1000
2000
3000
4000
5000
2
0
1000
2000
3000
4
0
6
0
1000
2000
3000
4000
5000
6000
7
0
8
0
1000
2000
3000
9
0
10
0
11
0
12
0
1000
2000
13
0
1000
2000
14
0
1000
17
0
1000
18
0
19
0
20
0
discharge
0
1000
2000
duration (days)
10+
7
4
1