JAMIA Open, 6(1), 2023, ooad002
https://doi.org/10.1093/jamiaopen/ooad002
Research and Applications

Research and Applications

Understanding comorbidities and health disparities related to COVID-19: a comprehensive study of 776 936 cases and 1 362 545 controls in the state of Indiana, USA

Nader Zidan[1], Vishal Dey[1], Katie Allen [2], John Price[2], Sarah Renee Zappone[2], Courtney Hebert[3], Titus Schleyer[2,4], and Xia Ning [1,3,5]

[1]Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio, USA, [2]Regenstrief Institute, Indianapolis, Indiana, USA, [3]Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio, USA, [4]Department of Medicine, School of Medicine, Indiana University, Indianapolis, Indiana, USA and [5]Translational Data Analytics Institute, The Ohio State University, Columbus, Ohio, USA

Corresponding Author: Xia Ning, PhD, 1800 Cannon Drive, 310C, Columbus, OH 43210, USA; ning.104@osu.edu
Nader Zidan and Vishal Dey are co-first authors.

Received 16 May 2022; Revised 4 December 2022; Editorial Decision 1 January 2023; Accepted 6 January 2023

ABSTRACT

Objective: To characterize COVID-19 patients in Indiana, United States, and to evaluate their demographics and comorbidities as risk factors to COVID-19 severity.
Materials and Methods: EHR data of 776 936 COVID-19 cases and 1 362 545 controls were collected from the COVID-19 Research Data Commons (CoRDaCo) in Indiana. Data regarding county population and per capita income were obtained from the US Census Bureau. Statistical analysis was conducted to determine the association of demographic and clinical variables with COVID-19 severity. Predictive analysis was conducted to evaluate the predictive power of CoRDaCo EHR data in determining COVID-19 severity.
Results: Chronic obstructive pulmonary disease, cardiovascular disease, and type 2 diabetes were found in 3.49%, 2.59%, and 4.76% of the COVID-19 patients, respectively. Such COVID-19 patients have significantly higher ICU admission rates of 10.23%, 14.33%, and 11.11%, respectively, compared to the entire COVID-19 patient population (1.94%). Furthermore, patients with these comorbidities have significantly higher mortality rates compared to the entire COVID-19 patient population. Health disparity analysis suggests potential health disparities among counties in Indiana. Predictive analysis achieved F1-scores of 0.8011 and 0.7072 for classifying COVID-19 cases versus controls and ICU versus non-ICU cases, respectively.
Discussion: Black population in Indiana was more adversely affected by COVID-19 than the White population. This is consistent to findings from existing studies. Our findings also indicate other health disparities in terms of demographic and economic factors.
Conclusion: This study characterizes the relationship between comorbidities and COVID-19 outcomes with respect to ICU admission across a large COVID-19 patient population in Indiana.

Key words: COVID-19, risk factor analysis, comorbidity, health disparity analysis, predictive analysis

Lay Summary

Our study has 3 goals: (1) to perform risk-factor analyses to relate pre-existing comorbidities with the risk of ICU admission for COVID-19 patients; (2) to perform health disparity analyses to examine the association between the prevalence of COVID-19 with epidemiological and socio-economic factors; (3) to perform predictive analyses to determine the predictive power of patient EHR data in predicting COVID-19 infection and need for ICU treatment. EHR data of 776 936 COVID-19 cases and 1 362 545 controls were collected from the COVID-19 Research Data Commons (CoRDaCo) in Indiana. From our risk-factor analysis, age and the presence of comorbidity were found to be associated with an increased risk of ICU admission and mortality. Furthermore, patients with comorbidities have significantly higher mortality rates compared to the entire COVID-19 patient population. Black patient population in Indiana was found to be more adversely affected by COVID-19 than the White patient population. Our analysis revealed potential health disparities that indicate differences in accessibility to healthcare and representation in CoRDaCo. The results from predictive analyses suggest that CoRDaCo EHR data are informative enough to predict a patient's risk of infection to COVID-19 and whether a COVID-19-infected patient needs ICU admission.

INTRODUCTION

On March 11, 2020, the World Health Organization declared the Coronavirus disease 2019 (COVID-19) a pandemic.[1, 2] COVID-19 is caused by the novel coronavirus severe acute respiratory syndrome coronavirus-2.[3, 4] As of October 17, 2022, there were more than 621 million reported cases and more than 6 million deaths globally.[5] The United States, with over 95 million cases reported and more than 1 million deaths, continues to be impacted by this virus.[6] The first case in Indiana was reported in March 2020.[7] As of October 17, 2022, there have been over 1.7 million reported cases and over 22 000 confirmed deaths attributed to COVID-19 in Indiana.[8] The most common symptoms of COVID-19 include fever, cough, and fatigue. Other symptoms include shortness of breath, headache, loss of taste and/or smell, diarrhea, and vomiting.[3, 9] Recent work has shown that the presence of chronic obstructive pulmonary disease (COPD), cardiovascular disease (CVD), hypertension, and diabetes mellitus may increase the risk of COVID-19 severity.[10, 11] In addition, older age, male sex, COPD, and diabetes mellitus are associated with increased risk of mortality.[12, 13] Understanding the relationship between comorbidities, and in general, health conditions, and COVID-19 patient outcomes is important to determine patient-specific prognoses for the disease.[14] While most existing work concentrates on COVID-19 patients across nations[15] or in specific hotspots,[16–20] not many have focused on the Midwest of the United States.[21–23] The Midwest has some of the highest prevalence of comorbidities, specifically hypertension, COPD, and diabetes.[24–26] Therefore, analysis of a large, comprehensive cohort of focused Midwest patients is important to better decipher the relationship between comorbidities and COVID-19 outcomes in this region. In this study, the EHR data of 776 936 COVID-19 patients (cases) and 1 362 545 patients with negative COVID-19 tests from the state of Indiana was examined.

Background and significance

A large volume of studies have examined the relationship between COVID-19 and comorbidities.[10, 27–33] This body of literature has allowed researchers to estimate the association of co-morbidities with major COVID-19 outcomes, such as hospitalization, ICU admission, and death. Most of the existing work used small cohorts and thus the results may not be representative. Although there are studies[16, 27, 34] with moderately large patient cohorts, they primarily considered patients who were critically ill and hospitalized with COVID-19. Specifically, Fried et al[27] presented the patient characteristics and identified clinical factors for 11 721 patients hospitalized for 7 days on average with COVID-19 across the United States. Although this study has one of the largest patient cohorts in the literature, the patient population is unevenly distributed across states with the Midwest being the most under-represented region (constituting only 4% of patients). Moreover, the conclusions from this study could have been dominated by the over-represented Northeast region, and might not be generalized to Indiana with good faith. In order to address this, Nanda et al[28] developed a well-characterized cohort of 1169 adult COVID-19 positive patients from the Midwest. However, the cohort size is quite small. Nonetheless, most of these studies did not consider cohorts in early or late stages of COVID-19. Meanwhile, the National COVID Cohort Collaborative (N3C)[15] established a very large and comprehensive cohort of over 3.3 million COVID-19 patients over 67 health care institutions. While there have been studies[35–37] that use the N3C data to study and characterize COVID-19, these studies did not focus on patients in Indiana. Our work is motivated to alleviate limitations of prior research in terms of sample size and geographic coverage, in addition to establishing findings in the state of Indiana using a relatively larger and more comprehensive cohort. Moreover, following robust data-driven methodologies, we also analyzed the predictive power of EHR data in predicting susceptibility and severity for a novel disease such as COVID-19. Furthermore, our work is one of the very few studies that involve such a large and comprehensive state-wide cohort of COVID-19 patients.

Furthermore, several studies[38–41] have examined the relationship between socio-economic factors and COVID-19. For example, Hatef et al[38] assessed the impact of neighborhood socio-economic factors on COVID-19 prevalence across 7 US states. They used American Community Survey (ACS) and US Census data to obtain zip code and demographic data (eg, age, race, and population size) to determine an Area Deprivation Index (ADI) which ranks neighborhoods by their socio-economic characteristics. They found that zip codes with a higher ADI (ie, more disadvantaged neighborhoods) had higher prevalence of COVID-19 compared to zip codes with a lower ADI for all states except Florida and Virginia. Abedi et al[39] reported the association between COVID-19 infection and mortality, race and economic inequality in 7 US states. They used data from USAFacts,[42] the US Census, COVID-19 data from each state, state population by race and ethnicity data, and mobility data extracted from Google. They found that counties with higher levels of education, income, and population were at greater risk of infection, while counties with lower income levels were at higher risk of mortality. However, both of these studies did not include Indiana as one of their 7 states.

Our study has 3 goals: (1) to perform risk-factor analyses to relate pre-existing comorbidities with the risk of ICU admission for

COVID-19 patients; (2) to perform health disparity analyses to examine the association between the prevalence of COVID-19 with epidemiological and socio-economic factors; and (3) to perform predictive analyses to determine the predictive power of patient EHR data in predicting COVID-19 infection and need for ICU treatment.

MATERIALS AND METHODS

EHR data were obtained from the COVID-19 Research Data Commons (CoRDaCo).[43] CoRDaCo was created and is managed by the Regenstrief Institute to efficiently collect, manage, and analyze datasets related to COVID-19 for research. CoRDaCo contains data from 3 sources: the Indiana Network for Patient Care (INPC) managed by the Indiana Health Information Exchange, and the clinical data warehouses of Indiana University Health and Eskenazi Health. We included data between January 1, 2018 and May 5, 2021. COVID-19 positive cases were confirmed via laboratory testing (eg, PCR test) or diagnoses represented by ICD codes. The controls had to meet all the following criteria: (1) they had at least 1 negative COVID-19 test result and no positive COVID-19 test result(s) and (2) they had at least 1 INPC encounter between 2018 and 2019. The latter criterion was required to eliminate patients for whom the only clinical variable was COVID-19 testing. In this study, we examined 12 comorbidities (Supplementary Table S1). These comorbidities were selected because: (1) they are prevalent in Indiana[44–46] and (2) they are demonstrated to have strong associations with COVID-19 severity in other states.[10, 27, 28] Comorbidities were identified using corresponding ICD-10 and ICD-9 codes (used in less than 1% of the records) as presented in Supplementary Table S1. Data analysis was conducted using Python version 3.8.3 over Indiana University's Carbonate High Performance Computing cluster with CPU Intel Xenon E5-2680 v3 and 256 GB of RAM.[47]

Collected data included demographics, diagnoses, laboratory results, medications, inpatient and outpatient encounter information including ICU status, and mortality for all patients.[43] In total, there were 776 759 COVID-19 positive patients and 1 362 527 COVID-19 negative patients between ages 0 and 110 years. Supplementary Figure S1 presents an overview of how the COVID-19 positive patient cohort was labeled and divided into groups based on comorbidities and ICU admission. A patient was given a label of "ICU" if the patient had an ICU encounter after March 6, 2020 (ie, the date of COVID-19 appearance in Indiana). Following such labeling scheme, in total, there were 15 084 cases and 35 743 controls labeled "ICU."[43] Patients were considered to have a comorbidity if the patient had a diagnosis of the comorbidity prior to their date of COVID-19 diagnosis or March 6, 2020 if they did not have a recorded COVID-19 diagnosis date. A more severe case of COVID-19 is considered as the one which resulted in an ICU admission. Socio-economic data, including per capita income and county population, were obtained from the US Census Bureau.[48, 49] Epidemiological data such as county-wise COPD prevalence was obtained from the CDC PLACES data portal.[46] Since per capita income data were not available with more granular place-based indicators (eg, zip codes), the epidemiological and socio-economic data were collected at the county level. The predictive analysis utilized demographic and clinical information obtained from a patient's EHR data such as age, gender, race, ethnicity, comorbidities, diagnoses, and clinical laboratory results (eg, White Blood Cell count).

Risk factor analysis

Continuous variables (eg, "age") were presented using their median, minimum, maximum and mean values, and their standard deviations. Categorical variables (eg, "gender", "race") were presented using counts and percentages. Note that the various race categories presented, such as Asian/Pacific Islander, were sourced from CoRDaCo without any modification. The Asian and Pacific Islander race categories were grouped together as 1 category in CoRDaCo because: (1) some health systems in Indiana supply the data with the 2 race groups combined, preventing disambiguation and, (2) when separated where feasible, the number of patients was limited. Therefore, the decision with the CoRDaCo race categories was to leave them combined. Statistical testing was conducted to determine the association of these variables with ICU admission. For continuous variables, two-sample *t*-test was applied to test their differences between ICU and non-ICU patients if the data were normally distributed; otherwise, Mann–Whitney test was applied. For categorical variables, χ^2 test was applied to test their association with ICU admissions, and Fisher's exact test was applied if data were limited. Confidence level was set to 95% for all statistical tests.

The *P*-values indicate whether there was a statistically significant difference between the ICU and non-ICU patients in terms of the variables (such as Age, Death Flag, Gender, etc.). Specifically, a *P*-value for Age variable indicates whether the difference between the mean age of ICU and non-ICU patients is statistically significant (ie, whether age has an effect on the ICU admission). This *P*-value is computed from the age of ICU and non-ICU patients. A *P*-value for Death Flag indicates whether ICU admission and death outcomes are associated. This *P*-value is computed from 4 counts: ICU patients alive, ICU patients deceased, non-ICU patients alive, and non-ICU patients deceased. A *P*-value for Gender indicates whether there is an association between ICU admission and gender categories. This *P*-value is computed from 6 counts: female ICU patients, male ICU patients, ICU patients with "Unknown" gender, female non-ICU patients, male non-ICU patients, and non-ICU patients with "Unknown" gender, A *P*-value for each race category indicates whether a given race is significantly associated with ICU admission. The *P*-value for a given race category is computed from 4 counts: ICU patients belonging to that race, ICU patients not belonging to that race, non-ICU patients belonging to that race, and non-ICU patients not belonging to that race.

Health disparity analysis

The CDC PLACES[46] dataset was used to obtain the county-wise COPD prevalences. The US Census[48] data were used to obtain the total county populations. The US Census Bureau[49] was used to obtain the ACS data on county per capita income. Patients were assigned to a county using their zip code. For zip codes that exist within multiple counties, the county which covered the most area of the zip code was used. The COVID-19 prevalence in CoRDaCo was computed as the percentage of the total county population that had COVID-19 in CoRDaCo. Correlations were computed pairwise between the total county population, county per capita income, county COPD prevalence, county COVID-19 prevalence in CoRDaCo, and county population in CoRDaCo. Heatmaps were generated in R to visualize the data for each county.

Predictive analysis

The objective of this predictive analysis was to evaluate whether CoRDaCo EHR data including demographic information,

comorbidities, and other clinical variables can predict COVID-19 susceptibility and severity. Constructing predictive models may reveal whether EHR data are informative enough to predict a patient's susceptibility to a novel disease such as COVID-19, and whether a COVID-19 infected patient needs ICU admission. Furthermore, some models can reveal important features which are most related to COVID-19 susceptibility and severity. In turn, these features can inform clinical decision-making for effective patient care. CoRDaCo consisted of 776 759 cases (ie, patients with COVID-19) and 1 362 527 controls (ie, patients without COVID-19). Cases were further grouped into ICU cases (ie, COVID-19 cases and admitted to ICU) and non-ICU cases (ie, COVID-19 cases but not admitted to ICU) based on the ICU labeling scheme described earlier in the Materials and Methods section. Following the objective, 2 classification tasks were designed: (1) to classify between cases and controls and (2) to classify between ICU cases and non-ICU cases. These tasks have prominent clinical significance—they can be used to determine whether a patient is susceptible to COVID-19, and whether a COVID-19 patient will need ICU admission in the future, respectively, based on the patient's EHR data.

A feature vector for each patient was constructed from EHR data before March 6, 2020 (ie, the first occurrence of COVID-19 in Indiana), including demographics, diagnoses, lab tests, and encounters information. Demographic information was represented as either numerical features (eg, age) or categorical features (eg, gender, race, ethnicity). Comorbidities were represented using binary features indicating the presence (1) or absence (0) of each of the 12 comorbidities presented in Supplementary Table S1. For diagnoses, the top-k ($k = 10$ or 50) most frequent ICD codes occurring among patients in each of the 3 groups (ie, ICU cases, non-ICU cases and controls) were selected, excluding ICD codes for the 12 comorbidities and COVID-19 diagnosis (ICD-10: U07.1). This resulted in 2 sets of features: (1) features consisting of 11 unique ICD codes (with $k = 10$), denoted as $\mathcal{D}_{11}$, and (2) features consisting of 58 unique ICD codes (with $k = 50$), denoted as $\mathcal{D}_{58}$. Frequencies of the ICD codes were used as the feature values. In addition, a total of 149 clinical lab tests were represented using either numerical or categorical features, depending on the result types. Categorical lab results such as "Unknown", "Inconclusive", and "test not known" were grouped into the "Unknown" category. For encounters, frequencies of unique outpatient, inpatient, emergency, and unknown encounters were used as features. Numerical features were standardized using z-score normalization, and categorical features were one-hot encoded.

For each task, a grid search was conducted over various hyperparameters (presented in Supplementary Tables S2–S4). Among all the patients, 1000 patients were randomly sampled from each of the 3 groups (ICU cases, non-ICU cases, and controls) independently. For each task, 5-fold cross validation was conducted. Logistic Regression (LR) and Gradient Boosting Tree (GDBT) were used to train the classifiers for each task. Precision, recall, accuracy, area under the receiver-operating curve (AUROC), and F1-score were used as the evaluation metrics. The model that produced the best F1-score was selected as the best model.

RESULTS

In this section, risk factor analyses with respect to COPD (Table 1), CVD (Supplementary Table S6) and type 2 diabetes (T2D) (Supplementary Table S8), health disparity analyses, and predictive analyses are presented. Additional results are presented in the Supplementary Materials. CoRDaCo data characteristics can be found in Allen et al.[43]

Risk factor analysis

Table 1 presents the characteristics of COVID-19 patients who had COPD. Among these 27 125 patients (3.49% of all COVID-19 patients), significantly more female were in both ICU and non-ICU populations; and Black/African American patients constituted a slightly higher proportion in the ICU compared to non-ICU. In addition, patterns were observed in patients with COPD and COVID-19 similar to those observed among all COVID-19 patients: age was significantly associated with ICU admission; death rates were significantly higher among ICU patients compared to non-ICU patients; gender was a risk factor for COVID-19 severity. When compared to all COVID-19 patients, there were some differences observed among the COVID-19 patients with COPD. First, ICU admission rate for COVID-19 patients with COPD was significantly higher (10.23% vs 1.94%[43]) indicating that COPD was a strong risk factor; second, the death rate among COVID-19 patients with COPD was significantly higher (8.22% vs 2.24%[43]).

Supplementary Table S6 presents the characteristics of COVID-19 patients with CVD (2.59% among all COVID-19 patients) and Supplementary Table S8 presents the characteristics of COVID-19 patients with T2D (4.76% among all COVID-19 patients). For both groups, we observed similar patterns as those observed among all COVID-19 patients[43] and also among COVID-19 patients with COPD. Age was significantly associated with ICU admission; death rates were significantly higher among ICU patients compared to non-ICU patients. However, when compared to all COVID-19 patients, those with COVID-19 and CVD exhibited significantly higher ICU admission rates (14.33% vs 1.94%[43]) and had much higher death rates (13.48% vs 2.24%[43]). Similarly, for COVID-19 patients with T2D when compared to all COVID-19 patients: ICU admission rates were significantly higher (11.11% vs 1.94%[43]) and death rates were significantly higher (9.16% vs 2.24%[43]).

When comparing COVID-19 patients with COPD, CVD, or T2D, it is noticeable that females constituted a higher proportion of ICU populations than males in all 3 comorbidities. In addition, Black/African American patients constituted a higher proportion in the ICU compared to non-ICU in all 3 comorbidities. However, there was a noticeable difference: gender was significantly associated with ICU admission in COVID-19 patients with COPD or T2D; but this was not the case in COVID-19 patients with CVD. Other interesting patterns were found among COVID-19 patients with COPD, CVD, or T2D when compared to the corresponding control patients with the same comorbidity. Such patterns are presented in the Supplementary Materials. Supplementary Tables S5, S7, and S9 present the characteristics of controls with COPD, CVD, and T2D, respectively.

COVID-19 patients with other comorbidities (as listed in Supplementary Table S1), when compared to all COVID-19 patients, had similar patterns as those observed when COVID-19 patients with COPD, CVD, or T2D were compared to all COVID-19 patients: the ICU admission rate was relatively higher, and the death rates were much higher. In fact, the highest ICU admission rate was 31.75% for COVID-19 patients with dependence on renal dialysis versus 1.94% ICU admission rate for all COVID-19 patients (P-value $<$ .05). Furthermore, COVID-19 patients with dependence on renal dialysis, cardiac arrest, and ARF had the 3 highest death rates of 24.32%, 20.56%, and 20.56%, respectively. Overall, these patterns suggest that these comorbidites are potential risk factors for COVID-19 severity and mortality.

Some patterns observed for COVID-19 patients with COPD and T2D were also observed for COVID-19 patients with comorbidities

Table 1. Characteristics of COVID-19 patients with COPD comorbidity

	N (%) Total (*N* = 27 125)	ICU (*N* = 2775)	Non-ICU (*N* = 24 350)	*P*-value[a]
Age; median (min–max), (Mean; SD)	58.66 (0.93–106.8) (54.83; 21.61)	69.55 (4.49–106.8) (67.54; 14.77)	56.95 (0.93–102.61) (53.38; 21.79)	<.05
Death Flag				<.05
Alive	24 895 (91.78)	1885 (67.93)	23 010 (94.50)	
Deceased	2230 (8.22)	890 (32.07)	1340 (5.50)	
Gender				<.05
Female	16 868 (62.19)	1546 (55.71)	15 322 (62.92)	
Male	10 255 (37.81)	1229 (44.29)	9026 (37.07)	
Unknown	2 (0.01)	0 (0.00)	2 (0.01)	
Race				
American Indian/Alaska Native	20 (0.07)	1 (0.04)	19 (0.08)	.72
Asian/Pacific Islander	205 (0.76)	15 (0.54)	190 (0.78)	.17
Black/African American	4082 (15.05)	447 (16.11)	3635 (14.93)	.10
Multiracial	125 (0.46)	13 (0.47)	112 (0.46)	.95
Other/unknown	593 (2.19)	15 (0.54)	578 (2.37)	<.05
White	22 100 (81.47)	2284 (82.31)	19 816 (81.38)	.23

Note: The *P*-value for Age variable indicates whether the difference between the mean age of ICU and non-ICU patients is statistically significant (ie, whether age has an effect on the ICU admission). The *P*-value for Death Flag indicates whether there exists an association between ICU admission and death outcomes. The *P*-value for Gender indicates whether there is an association between ICU admission and gender categories. The *P*-value for each race category indicates whether a given race is significantly associated with ICU admission.

COPD: chronic obstructive pulmonary disease; ICU: intensive care unit.

[a]*P*-value < .05 was used to determine statistical significance.

such as HTN and other cardiac arrhythmias: both age and gender were significantly associated with ICU admission (*P*-value < .05). However, for patients with ARF, dependence on renal dialysis, and cardiac arrest, neither age nor gender was found to be significantly different across ICU and non-ICU patients. For patients with acute myocardial infarction, VTE, and PE, age was found to be significantly higher across ICU and non-ICU patients, but gender was not. Furthermore, there were some noticeable patterns with respect to race. For COVID-19 patients with any of our considered comorbidities, Black/African American patients constituted a larger proportion of ICU-admitted patient population compared to that of the non-ICU admitted patient population; the population proportion of White patients in the ICU was lower than that in the non-ICU except for patients with COPD.

Overall, COVID-19 patients with any of our considered comorbidities had higher death rates than controls (ie, patients without COVID-19) with that comorbidity; ICU death rate was also higher for COVID-19 patients with comorbidities than for controls with that comorbidity. Furthermore, for patients with comorbidities related to heart and kidney diseases such as CVD, ARF, cerebral infarction, acute myocardial infarction, and cardiac arrest, the mean age of COVID-19 patients was higher than that among controls. Generally, for controls with a comorbidity, ICU patients tended to be older than non-ICU patients. However, this was not the case for some comorbidities: dependence on renal dialysis, and cardiac arrest. A similar pattern was observed for COVID-19 patients with either of these 2 comorbidities, where age was not found to be significantly associated with ICU admission.

Health disparity analysis

We observed that the prevalence of HTN, T2D, and COPD in CoRDaCo (ie, 9.52%, 4.76%, and 3.49%) is lower than nationwide (ie, 45.40%,[50] 10.50%,[25] and 6.40%[51]) and in Indiana (ie, 34.80%,[52] 12.50%,[45] and 8.30%[52]). Figure 1a–e present the heatmaps of Indiana that display the county-wise crude COPD prevalence,[46] COVID-19 prevalence[48] (ie, count of COVID-19 patients in a county in CoRDaCo over the total county population), per capita income,[49] the COVID-19 patient population in CoRDaCo, and the count of COVID-19 patients with ICU admission, respectively. The crude COPD prevalence as in Figure 1a shows that some counties (eg, Crawford) have COPD prevalence as high as 13% and others (eg, Hamilton) as low as 6%. Unfortunately, counties with higher COPD prevalence are not always proportionally represented in CoRDaCo (eg, Crawford in Figure 1a and b). As a matter of fact, counties with higher COPD prevalence tend to be underrepresented in CoRDaCo (eg, LaGrange in Figure 1a and b), leading to lower COPD prevalence in CoRDaCo.

As Figure 1b shows, in CoRDaCo, the COVID-19 patient population in each county does not reflect the Indiana population in that county proportionally. For some of the counties (eg, Clinton, Shelby), about 15% of the population was recorded as infected with COVID-19 in CoRDaCo, but in other counties (eg, Franklin, Clay), the percentage is 7% and as low as 4%. Comparing Figure 1a–c, we observed that counties with higher per capita income tend to be better represented in CoRDaCo (eg, Boone, $44 712.00, 10.22%), and tend to have low COPD prevalence (eg, 6.60%). In addition, the heatmaps of the COVID-19 patient population (Figure 1d) and the ICU admitted COVID-19 patient population (Figure 1e) display similar intensities across counties, indicating a high correlation between the 2 populations.

In order to better understand the associations among these epidemiological and socio-economic factors, we further studied the pairwise correlations among them. Figure 2 presents the pairwise Pearson correlations among the entire county population of each county (County Pop), count of COVID-19 positive patients in CoRDaCo (Patient Count), count of COVID-19 positive patients in CoRDaCo labeled ICU (ICU Count), per capita income (Income), COVID-19 prevalence (COVID%), and COPD prevalence

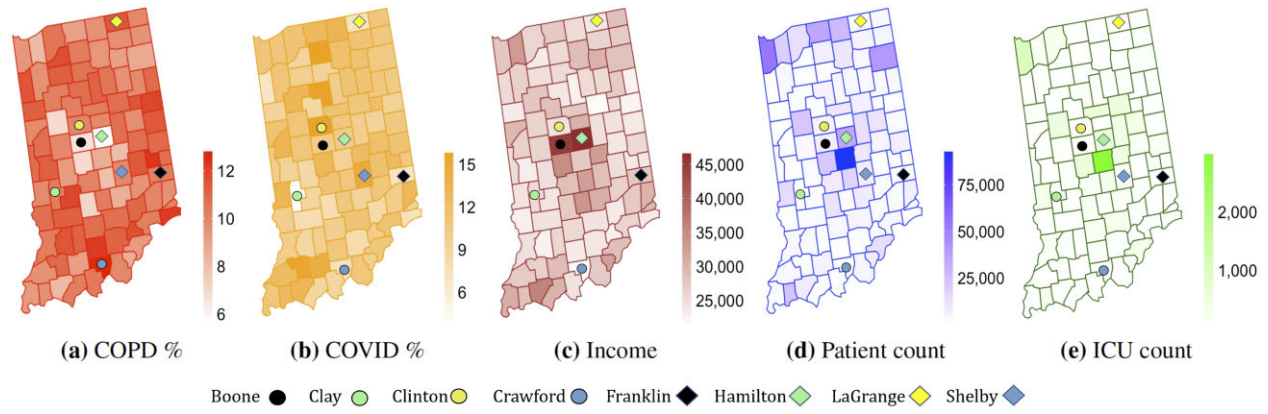

Figure 1. County-wise heatmap of Indiana displaying various epidemiological and socio-economic factors.

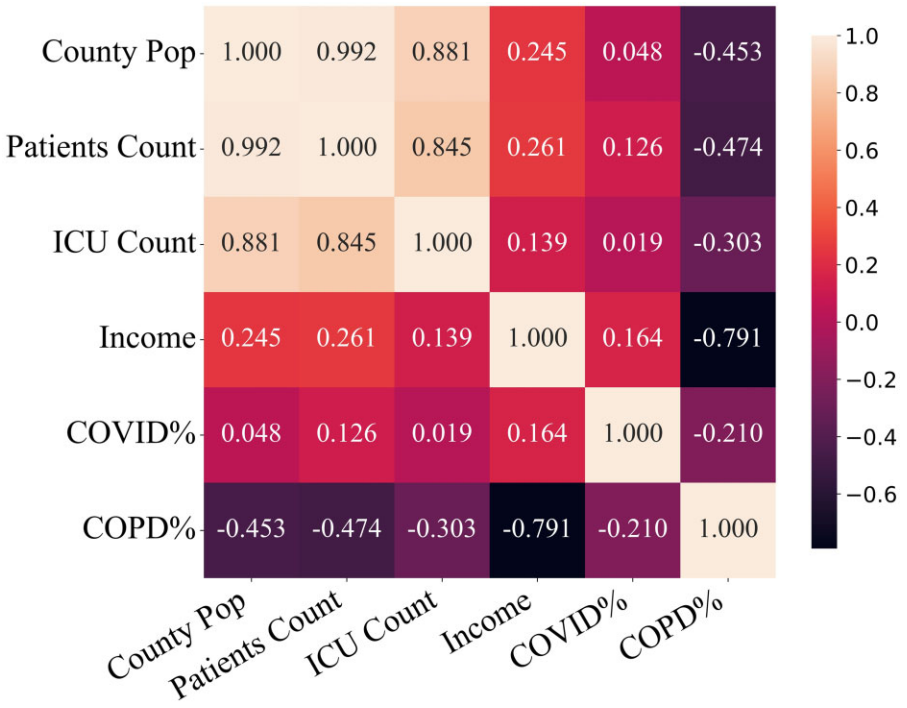

Figure 2. Correlation plot between various epidemiological and socio-economic factors.

(COPD%). The plot demonstrates some strong patterns: the correlation among the county population, patients count, and ICU count are all high and positive; the correlation between income and county population is moderate; the correlation between county population and COVID% is negligible; finally, the correlation between per capita income and COPD prevalence is high and negative. We further expand on these patterns in the discussion.

Predictive analysis

In order to evaluate whether CoRDaCo EHR data can effectively be leveraged to predict COVID-19 susceptibility, predictive models are constructed to classify cases versus controls. Table 2 presents the overall performance achieved by all the best models for classifying cases versus controls. Note that the best model is the one that achieves the best F1-score averaged over 5-folds. Clearly, the set of hyperparameters that yields the best F1-score averaged over 5-folds is the best configuration. We reported the best GDBT configuration for each task in Supplementary Tables S3 and S4. The reported performance metrics in Tables 2 and 3 are obtained from the best configuration for each model. Note that the performance of each model in metrics other than F1-score does not necessarily correspond to the optimal performance in those metrics. The best LR model achieved an F1-score of 0.8000 using $\mathcal{D}_{11}$. Moreover, the LR models achieved comparable performance across other metrics using $\mathcal{D}_{11}$ and $\mathcal{D}_{58}$. The best GDBT model achieved an F1-score of 0.8011 using $\mathcal{D}_{58}$, slightly better than that from LR. The GDBT models using $\mathcal{D}_{11}$ and $\mathcal{D}_{58}$ had similar performance in terms of precision, F1-score, and accuracy, marginally better than or comparable to those from best LR model. Clearly, the additional diagnosis information available in $\mathcal{D}_{58}$ provides minimal boost to the overall performance. This might indicate that the diagnoses features are not quite informative for distinguishing cases and controls. This aligns with our analysis of the feature importance as discussed below. Supplementary Table S10

Table 2. Overall performance for classifying cases versus controls

Model	Features	Precision	Recall	F1-score	Accuracy	AUROC
LR	$\mathcal{D}_{11}$	0.6667	1.0000	0.8000	0.6667	0.6541
	$\mathcal{D}_{58}$	0.6666	0.9995	0.7997	0.6663	0.6576
GDBT	$\mathcal{D}_{11}$	0.6703	0.9955	0.8010	0.6703	0.6698
	$\mathcal{D}_{58}$	0.6687	0.9990	0.8011	0.6693	0.6717

Table 3. Overall performance for classifying ICU versus non-ICU cases

Model	Features	Precision	Recall	F1-Score	Accuracy	AUROC
LR	$\mathcal{D}_{11}$	0.5489	0.8995	0.6807	0.5790	0.6803
	$\mathcal{D}_{58}$	0.5702	0.8450	0.6790	0.6030	0.6706
GDBT	$\mathcal{D}_{11}$	0.5944	0.8754	0.7072	0.6385	0.7159
	$\mathcal{D}_{58}$	0.5690	0.9298	0.7057	0.6135	0.7129

presents the top-10 most important features for the best GDBT model with $\mathcal{D}_{58}$. We observed that the encounter feature "Outpatient count", the most important feature, has a very high importance compared to other features. Furthermore, we observed that the encounter feature "Emergency count" and demographic feature "Age" are also among the top most important features. From the additional diagnoses in $\mathcal{D}_{58}$, "Chronic kidney disease unspecified," "Fatigue," "Tobacco use," and "Anemia" are the only ones that appears among the top 10 most important features; however, their importance is low relative to the encounter and demographic features.

In order to evaluate whether CoRDaCo EHR data can effectively predict COVID-19 severity, predictive models are constructed to classify ICU vs. non-ICU cases. Table 3 presents the overall performance achieved by all the best models for classifying ICU versus non-ICU cases. The best LR model achieved an F1-score of 0.6807 using $\mathcal{D}_{11}$. with better recall and AUROC; however, with worse precision and accuracy than the LR model with $\mathcal{D}_{58}$. The best GDBT model achieved an F1-score of 0.7072 using $\mathcal{D}_{11}$, with relatively better precision, accuracy and AUROC than the GDBT model with $\mathcal{D}_{58}$. Overall, the best GDBT model outperformed the best LR model by 3.9% in terms of F1-score (0.7072 vs 0.6807). Supplementary Table S11 presents the top 10 most important features for the best GDBT model with $\mathcal{D}_{11}$. We observed that the demographic feature "Age," encounter features "Outpatient count" and "Inpatient count," and the comorbidity feature "CVD" are among the top most important features. Similar to the cases versus controls task, we observed that age and encounter counts were more important features than diagnoses.

DISCUSSION

The patterns observed among COVID-19 patients in our data have also been observed in other similar studies. For example, Hoang and Anh[10] found that COPD, CVD, cerebrovascular accident, hypertension, diabetes, and malignancy are strongly associated with increased COVID-19 severity as well as ICU admission. In our data, we also observed that COVID-19 patients with COPD, CVD, HTN, and T2D had higher mortality rates compared to all COVID-19 patients. Our conclusion that the presence of CVD in COVID-19 patients is strongly associated with increased risk of mortality, is also supported by Fried et al.[27] Furthermore, we observed that older age was significantly associated with ICU admission in patients with COVID-19 and T2D, which Lei et al[32] observed as well in their study of 288 adult patients with COVID-19 from China. These observations bolster the support for the conclusions drawn from our results. Below, we discuss and interpret our results from each analysis to reveal some interesting patterns. Finally, we discuss the strengths and limitations of our work.

Risk factor analysis

We observed some patterns in the ICU patient proportions which are indicative of health disparities. First, the Black/African American ICU population proportion is larger than the non-ICU population proportion among COVID-19 patients for all comorbidities. While the same pattern appears for all comorbidities except COPD among the controls, the difference between the proportions is not significant except for a few comorbidities (such as COPD, CVD, VTE, ARF, and acute myocardial infarction). Second, the White ICU population proportion is always lower among COVID-19 patients with comorbidities except COPD. However, among controls with comorbidities, the White ICU population proportion is lower in all comorbidities except HTN, T2D, and COPD. These patterns indicate that it is more likely to observe Black/African American patients with comorbidities to be admitted into ICU with COVID-19 than without COVID-19. Conversely, it is less likely to observe White patients with comorbidities to be admitted into ICU with COVID-19 than without. Thus, we infer that Black/African American patients with comorbidities may be more adversely affected by COVID-19 and are at a greater risk of ICU admission, thereby indicating a health disparity. Dixon et al[53] also found evidence of health disparity in the Black/African American community, notably that the COVID-19 burden on the Black/African American community was larger when compared to the COVID-19 burden on the White community.

Health disparity analysis

The patterns observed in Figures 1 and 2 also imply other broader health disparities in terms of demographic and economic factors (eg, income). As reported in the previous section, Figure 1b illustrates the disproportionate representation of county-wise COVID-19 patient population in CoRDaCo. While we confirmed that CoRDaCo well represents the overall Indiana population, the significant difference in the COVID-19 population representation may be due to 2 reasons: (1) some counties were hit by COVID-19 much worse than others; or (2) health data were not proportionally collected from all the counties. The latter is likely due to the following: healthcare may not have been equally accessible in all counties, healthcare may have been accessible but not affordable for residents of certain counties, and healthcare networks may not have been IHIE members, meaning that the patient data were not retrieved. Regardless, both of these 2 possible reasons indicate potential health disparities among all the counties. To better reveal such disparities, we examine more closely the patterns observed in Figure 2.

First, CoRDaCo well represents the entire population in all the counties in Indiana (correlation between COVID-19 positive patients and population is 0.992, P-value = 7.13e-82). In addition, the correlation between the count of COVID-19 patients in a county and the count of ICU-admitted COVID-19 patients from a county is very high (0.845), indicating that patterns found among the larger county population also translate to similar patterns among the ICU population. Note that this is consistent to the observations from Figures 1d and e. In addition, per capita income is moderately

positively associated with the population (0.245), indicating that more populated counties tend to be wealthier. In spite of the fact that more populated counties tend to be wealthier and thus having better access to medical facilities, such counties may not be more resistant to COVID-19 compared to other counties. This is evident from the finding that COVID-19 prevalence is not associated with the county population (the correlation between COVID-19 prevalence and county population is 0.048, close to 0). Thus, populated counties may have high or low prevalence of COVID-19. COPD prevalence is highly negatively associated with per capita income (–0.791, *P*-value = 6.6e-21), indicating strong health disparities. The wealthier counties may be able to afford better and more accessible healthcare, increasing their resilience to certain comorbidities, particularly chronic diseases. COPD prevalence is negatively associated with COVID-19 prevalence (–0.210) and total COVID-19 patients (–0.474) geographically, explaining why COPD prevalence in CoRDaCo is lower than that in the entire Indiana population. Note that this is not conflicting with that COPD is a risk factor of COVID-19 overall among all the population together (Table 1). This is also evidenced by the negative relationship between ICU county population and COPD prevalence (–0.303) and the lack of association between ICU county population and COVID prevalence (0.019).

Predictive analysis

Additionally, we evaluated the predictive power of the collected CoRDaCo EHR data to classify cases versus controls, and ICU versus non-ICU cases. We should point out that all of our models are tuned in terms of F1 performance. Hence, the reported performance in other metrics is from the model that achieved the best F1. In doing so, with best F1, high sensitivity is achieved. This is actually desired since we believe that missing a COVID-19 case or an ICU case (false negative) could have significant clinical implications. Overall, GDBT demonstrated empirically better or competitive performance as LR for both tasks. Furthermore, GDBT has the additional advantage of interpretability in that the relative feature importance is readily available. On the other hand, the coefficients associated with each feature in LR cannot readily determine the relative importance of features.

Our predictive analyses indicate that in addition to demographic information, encounter and comorbidity information available in the EHR data are crucial in both classification tasks. Specifically, for classifying cases versus controls, we observed that the outpatient and emergency encounter counts are among the top most important features. This maybe because less healthy patients are more likely to seek frequent medical care and emergency treatment. Such patients may also be more susceptible to diseases such as COVID-19 than relatively healthier ones. For classifying ICU versus non-ICU cases, we also observed that outpatient, inpatient, and emergency encounter counts are among the top most important features. Additionally, for this task, we observed that the age has relatively higher importance than other demographic information. This aligns with our findings from the risk factor analysis indicating that elder people are more likely to be admitted to ICU. Moreover, we found that the presence of CVD in a patient is an important feature to distinguish ICU versus non-ICU cases. This also complements our findings from the risk-factor analysis that the CVD comorbidity is a strong risk factor for COVID-19 patients. Overall, the diagnoses and clinical labs features did not seem to be very informative in either classification task. This could be due to the limited available information on diagnoses and clinical labs in CoRDaCo. We will discuss this limitation in detail later in this section.

Strengths and limitations

Our study has a few strengths. First, the size of our populations for both COVID-19 positive and COVID-19 negative is much larger than most studies.[10, 16, 27–34] This empowers us with the unique ability to examine patterns for a large population group within 1 specific demographic region. Additionally, unlike the aforementioned studies that consider clinical diagnoses over a relatively short time interval, our study includes all diagnoses and encounters of COVID-19 positive patients since 2018. Hence, our study can better capture comorbidities and other potential risk factors that might later lead to serious COVID-19 cases. Our study also includes outpatient and emergency encounters, whereas most studies[17, 22, 54, 55] only consider inpatient encounters. Thus, our study involves diverse clinical settings, and conclusions from our analysis can be better generalized.

However, our study also has some limitations. A drawback of assigning an ICU label to patients if they had an ICU encounter after March 6, 2020 is that the reason for ICU admission may be unrelated to COVID-19. Unfortunately, from the EHR data, the true reason was unavailable. We acknowledge that there are other possible ways of assigning ICU labels apart from the one used in this work. An alternative way of ICU labeling patients with COVID-19 would use the date of the COVID-19 test and ICU admission, and determine if an ICU admission "closely" follows a positive test. However, determining how "close" an ICU admission must be to a positive test may pose an additional challenge, and it may still be the case that the ICU admission is unrelated to COVID-19. While our data collection included inpatient observations such as vitals, these data were limited to less than 1% of the total COVID-19 positive population in our dataset. Medication data availability is a limitation of the INPC. In addition, for many clinical labs, the number of patients with data available was low. Thus, we did not study medications and clinical labs, which have been demonstrated highly useful for understanding COVID-19 patients.[13, 56, 57] Furthermore, of the 776 936 COVID-19 positive patients, diagnoses records were only present for 429 337 (55.26%) patients which may have resulted in some uncaptured patterns. As a result, some bias may be introduced in the sample with only 55.26% having clinical features available. Those who are healthier may not have much available diagnoses data, leading to findings that are more reflective of the less healthy demographic of Indiana. Furthermore, those who may have limited access to healthcare may not have clinical data available in CoRDaCo, meaning that the patterns we observed may not be inclusive of those who live in areas with limited access, such as rural or minority communities. Finally, the data gathered in CoRDaCo were gathered for clinical practice, not for research.[43] Therefore, the data do not have the highest levels of consistency, which could have resulted in uncaptured patterns as well. In future work, we intend to increase the breadth of data available to each patient. Increasing the data available per patient will allow our analysis to include more patients, which may strengthen current findings or reveal new patterns. Furthermore, we will also expand the parameters of what defines comorbidities. In this work, to define a comorbidity, we solely relied on ICD codes from diagnoses data. By incorporating other clinical factors such as vitals, medications, and laboratory test data in defining comorbidities, we may be able to better explore clinical characteristics of patients with that comorbidity. In the future, vaccine information will be available

for many patients which would allow for further exploration of health disparities by examining who and how early someone may have had access to a vaccine.

CONCLUSION

This study provides insight and adds to the current volume of literature that characterizes the relationship between patients with COVID-19 and comorbidities and disease severity. In this study, we examined the relationship between demographic and comorbidity presence as risk factors of COVID-19 severity. Overall, comorbidity presence was linked to an increased risk of ICU admission as well as mortality. For many comorbidities, age, gender, or both were strongly associated with ICU admission. During our analysis, we noticed patterns that were indicative of health disparities. We concluded that the Black/African American community in Indiana was more adversely affected by COVID-19 compared to the White community. Furthermore, our analysis revealed health disparities that indicate differences in accessibility to healthcare and representation in large datasets such as CoRDaCo.

FUNDING

This work was supported by the National Library of Medicine under Grant Number 3R01LM012605-03S1 and 1R21LM013678-01, the Indiana Clinical and Translational Sciences Institute (funded in part by Award Number UL1TR002529 from the National Institutes of Health, National Center for Advancing Translational Sciences) Clinical and Translational Sciences Award, and the Lilly Endowment, Inc. Physician Scientist Initiative. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

AUTHOR CONTRIBUTIONS

XN conceived the research and supervised NZ and VD; XN and TS obtained funding for the research; KA, JP, SRZ, CH, and TS provided substantial clinical background and insights; NZ, VD, and XN conducted the research, including data processing, methodology design, implementation, and analysis; NZ and VD drafted the original manuscript; NZ, VD, and XN conducted the manuscript editing; KA, CH, and TS reviewed the manuscript and provided constructive suggestions and feedback. All authors reviewed the final manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *JAMIA Open* online.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The data underlying this article were accessed from multiple sources. EHR data were collected from the COVID-19 Research Data Commons (CoRDaCo) which are created and managed by the Regenstrief Institute. The data can be accessed upon request at https://www.ridata.org/ or via email to askrds@regenstrief.org. Per-capita income data were obtained from the US Census Bureau, and are publicly available at https://data.census.gov/cedsci/. Epidemiological data were obtained from the CDC PLACES data portal and are publicly available at https://chronicdata.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-County-Data-20/swc5-untb/data.

REFERENCES

1. Muralidar S, Ambi SV, Sekaran S, Krishnan UM. The emergence of COVID-19 as a global pandemic: understanding the epidemiology, immune response and potential therapeutic targets of SARS-CoV-2. *Biochimie* 2020; 179: 85–100.
2. WHO. Listings of WHO's response to COVID-19. https://www.who.int/news/item/29-06-2020-covidtimeline, 2020. Accessed September 17, 2021.
3. Rothan HA, Byrareddy SN. The epidemiology and pathogenesis of coronavirus disease (covid-19) outbreak. *J Autoimmun* 2020; 109: 102433.
4. Joost Wiersinga W, Prescott HC. What is covid-19? *JAMA* 2020; 324 (8): 816.
5. WHO. Who coronavirus (covid-19) dashboard. https://covid19.who.int/, 2022. Accessed October 17, 2022.
6. CDC. Vaccinations > variants. https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/past-reports/09172021.html, 2021. Accessed September 17, 2021.
7. IDOH. State health department confirms 1st case of COVID-19 in Hoosier with recent travel. https://events.in.gov/event/state-health-department-confirms-1st-case-of-covid-19–in-hoosier-with-recent-travel/, 2020. Accessed September 18, 2021.
8. CDC. United States covid-19 cases and deaths by state. https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36/data, 2021. Accessed March 29, 2022.
9. CDC. Symptoms of COVID-19. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html, 2021. Accessed September 18, 2021.
10. Hoang T, Anh TTT. Comparison of comorbidities in relation to critical conditions among coronavirus disease 2019 patients: a network meta-analysis. *Infect Chemother* 2021; 53 (1): 13–28.
11. Li X, Xu S, Yu M, *et al.* Risk factors for severity and mortality in adult covid-19 inpatients in Wuhan. *J Allergy Clin Immunol* 2020; 146 (1): 110–8.
12. Iaccarino G, Grassi G, Borghi C, Ferri C, Salvetti M, Volpe M. Age and multimorbidity predict death among COVID-19 patients: results of the SARS-RAS study of the Italian society of hypertension. *Hypertension* 2020; 76: 366–72.
13. Gupta S, Hayek SS, Wang W, *et al.*; STOP-COVID Investigators. Factors associated with death in critically ill patients with coronavirus disease 2019 in the us. *JAMA Intern Med* 2020; 180 (11): 1436–47.
14. Cho SI, Yoon S, Lee HJ. Impact of comorbidity burden on mortality in patients with COVID-19 using the Korean health insurance database. *Sci Rep* 2021; 11 (1): 6375.
15. Haendel MA, Chute CG, Bennett TD, *et al.*; N3C Consortium. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc* 2021; 28 (3): 427–43.
16. Shah P, Owens J, Franklin J, *et al.* Demographics, comorbidities and outcomes in hospitalized covid-19 patients in rural southwest Georgia. *Ann Med* 2020; 52 (7): 354–60.
17. Myers LC, Parodi SM, Escobar GJ, Liu VX. Characteristics of hospitalized adults with COVID-19 in an integrated health care system in California. *JAMA* 2020; 323 (21): 2195–8.
18. Filardo TD, Khan MR, Krawczyk N, *et al.* Comorbidity and clinical factors associated with COVID-19 critical illness and mortality at a large public hospital in New York city in the early phase of the pandemic (March–April 2020). *PLoS One* 2020; 15 (11): e0242760.
19. Vahidy FS, Drews AL, Masud FN, *et al.* Characteristics and outcomes of COVID-19 patients during initial peak and resurgence in the Houston metropolitan area. *JAMA* 2020; 324 (10): 998–1000.

20. Wadhera RK, Wadhera P, Gaba P, *et al.* Variation in COVID-19 hospitalizations and deaths across New York city boroughs. *JAMA* 2020; 323 (21): 2192–5.
21. Miller J, Fadel RA, Tang A, *et al.* The impact of sociodemographic factors, comorbidities, and physiologic responses on 30-day mortality in coronavirus disease 2019 (COVID-19) patients in metropolitan Detroit. *Clin Infect Dis* 2021; 72 (11): E704–10.
22. Twigg HL, Khan SH, Perkins AJ, *et al.* Mortality rates in a diverse cohort of mechanically ventilated patients with novel coronavirus in the urban Midwest. *Crit Care Explor* 2020; 2 (8): e0187.
23. Harrison SL, Fazio-Eynullayeva E, Lane DA, Underhill P, Lip GY. Comorbidities associated with mortality in 31,461 adults with COVID-19 in the United States: a federated electronic medical record analysis. *PLoS Med* 2020; 17 (9): e1003321.
24. Newman D, Tong M, Levine E, Kishore S. Prevalence of multiple chronic conditions by U.S. state and territory, 2017. *PLoS One* 2020; 15 (5): e0232346.
25. Centers for Disease Control and Prevention. National diabetes statistics report, 2020. https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf. Accessed October 13, 2021.
26. Centers for Disease Control and Prevention. Facts about hypertension — cdc.gov. https://www.cdc.gov/bloodpressure/facts.htm, 2021. Accessed October 18, 2021.
27. Fried MW, Crawford JM, Mospan AR, *et al.* Patient characteristics and outcomes of 11 721 patients with coronavirus disease 2019 (COVID-19) hospitalized across the United States. *Clin Infect Dis* 2021; 72 (10): e558–65.
28. Nanda S, Toussaint L, Vincent A, *et al.* A midwest COVID-19 cohort for the evaluation of multimorbidity and adverse outcomes from COVID-19. *J Prim Care Community Health* 2021; 12: 215013272110109.
29. He F, Quan Y, Lei M, *et al.* Clinical features and risk factors for ICU admission in COVID-19 patients with cardiovascular diseases. *Aging Dis* 2020; 11 (4): 763–9.
30. Al-Salameh A, Lanoix JP, Bennis Y, *et al.* Characteristics and outcomes of covid-19 in hospitalized patients with and without diabetes. *Diabetes Metab Res Rev* 2021; 37 (3): e3388.
31. Lim MA, Pranata R, Huang I, Yonas E, Soeroto AY, Supriyadi R. Multiorgan failure with emphasis on acute kidney injury and severity of covid-19: systematic review and meta-analysis. *Can J Kidney Health Dis* 2020; 7: 205435812093857.
32. Lei M, Lin K, Pi Y, *et al.* Clinical features and risk factors of ICU admission for COVID-19 patients with diabetes. *J Diabetes Res* 2020; 2020: 5237840.
33. Sanyaolu A, Okorie C, Marinkovic A, *et al.* Comorbidity and its impact on patients with COVID-19. *SN Compr Clin Med* 2020; 2 (8): 1069–76.
34. Richardson S, Hirsch JS, Narasimhan M, *et al.*; the Northwell COVID-19 Research Consortium. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York city area. *JAMA* 2020; 323 (20): 2052–9.
35. Bennett TD, Moffitt RA, Hajagos JG, *et al.*; National COVID Cohort Collaborative (N3C) Consortium. Clinical characterization and prediction of clinical severity of SARS-CoV-2 infection among US adults using data from the US National COVID Cohort Collaborative. *JAMA Netw Open* 2021; 4 (7): e2116901.
36. Martin B, DeWitt PE, Russell S, *et al.* Characteristics, outcomes, and severity risk factors associated with SARS-CoV-2 infection among children in the US National COVID Cohort Collaborative. *JAMA Netw Open* 2022; 5 (2): e2143151.
37. Pfaff ER, Girvin AT, Bennett TD, *et al.* Who has long-COVID? A big data approach [published online ahead of print 2021]. medRxiv. https://www.medrxiv.org/content/10.1101/2021.10.18.21265168v1. Accessed October 4, 2022.
38. Hatef E, Chang HY, Kitchen C, Weiner JP, Kharrazi H. Assessing the impact of neighborhood socioeconomic characteristics on COVID-19 prevalence across seven states in the United States. *Front Public Health* 2020; 8: 9.
39. Abedi V, Olulana O, Avula V, *et al.* Racial, economic, and health inequality and COVID-19 infection in the United States. *J Racial Ethn Health Disparities* 2021;8 (3): 732–42.
40. Lo CH, Nguyen LH, Drew DA, *et al.* Race, ethnicity, community-level socioeconomic factors, and risk of COVID-19 in the United States and the United Kingdom. *EClinicalMedicine* 2021; 38: 101029.
41. Patel AP, Paranjpe MD, Kathiresan NP, Rivas MA, Khera AV. Race, socioeconomic deprivation, and hospitalization for COVID-19 in English participants of a national biobank. *Int J Equity Health* 2020; 19 (1): 114.
42. USAFacts. Coronavirus locations: COVID-19 map by county and state. https://usafacts.org/visualizations/coronavirus-covid-19-spread-map. Accessed October 4, 2022.
43. Allen K, Zidan N, Dey V, *et al.* Resource profile: the Regenstrief Institute COVID-19 research data commons (CoRDaCo). medRxiv 2021. https://www.medrxiv.org/content/10.1101/2021.12.17.21267942v1. Accessed January 5, 2022.
44. Indiana State Department of Health. Healthy people 2020 update: COPD. https://www.in.gov/health/cdpc/files/HP2020_COPD_FINAL5.pdf. Accessed October 4, 2022.
45. Indiana State Department of Health. Indiana diabetes strategic plan 2020–2026. https://www.in.gov/health/files/db.pdf. Accessed October 17, 2021.
46. Centers for Disease Control and Prevention. Places: local data for better health, county data 2020 release—chronic disease and health promotion data & indicators. https://chronicdata.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-County-Data-20/swc5-untb/data. Accessed October 12, 2021.
47. Indiana University Information Technology Services. About carbonate at Indiana University. https://kb.iu.edu/d/aolp. Accessed September 25, 2021.
48. United States Census Bureau. County population totals: 2010–2019. https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html. Accessed October 13, 2021.
49. United States Census Bureau. Census – table results. https://data.census.gov/cedsci/. Accessed October 7, 2021.
50. Ostchega Y, Fryar CD, Nwankwo T, Nguyen DT. Hypertension prevalence among adults aged 18 and over: United States, 2017–2018. https://www.cdc.gov/nchs/products/databriefs/db364.htm, 2020. Accessed October 13, 2021.
51. Biener AI, Decker SL, Rohde F. Prevalence and treatment of chronic obstructive pulmonary disease (COPD) in the United States. *JAMA* 2019; 322 (7): 602.
52. Centers for Disease Control, Prevention, National Center for Chronic Disease Prevention. Health Promotion, and Division of Population Health. *BRFSS* Prevalence & Trends Data. https://www.cdc.gov/brfss/brfssprevalence/, 2015. Accessed October 13, 2021.
53. Dixon BE, Grannis SJ, Lembcke LR, Valvi N, Roberts AR, Embi PJ. The synchronicity of covid-19 disparities: Statewide epidemiologic trends in sars-cov-2 morbidity, hospitalization, and mortality among racial minorities and in rural America. *PLoS One* 2021; 16 (7): e0255063.
54. Roth GA, Emmons-Bell S, Alger HM, *et al.* Trends in patient characteristics and COVID-19 in-hospital mortality in the United States during the COVID-19 pandemic. *JAMA Netw Open* 2021; 4 (5): e218828.
55. Finelli L, Gupta V, Petigara T, Yu K, Bauer KA, Puzniak LA. Mortality among US patients hospitalized with SARS-CoV-2 infection in 2020. *JAMA Netw Open* 2021; 4 (4): e216556.
56. Chen Y, Yang D, Cheng B, *et al.* Clinical characteristics and outcomes of patients with diabetes and COVID-19 in association with glucose-lowering medication. *Diabetes Care* 2020; 43 (7): 1399–407.
57. Sterne JA, Murthy S, Diaz JV, *et al.*; WHO Rapid Evidence Appraisal for COVID-19 Therapies (REACT) Working Group. Association between administration of systemic corticosteroids and mortality among critically ill patients with covid-19: a meta-analysis. *JAMA* 2020; 324 (13): 1330–41.