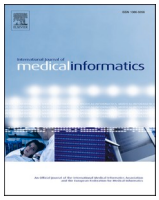
International Journal of Medical Informatics 166 (2022) 104842

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Using EHR data to identify coronavirus infections in hospitalized patients: Impact of case definitions on disease surveillance

Ann Marie Navar [a,*], Irene Cosmatos [b], Stacey Purinton [c], Janet L. Ramsey [b], Robert J. Taylor [c], Rachel E. Sobel [b], Ginger Barlow [b], Gretchen S. Dieck [b], Michael L. Bulgrein [b], Eric D. Peterson [a]

[a] University of Texas Southwestern Medical Center, Dallas, TX, United States
[b] UBC, Philadelphia, PA, United States
[c] Cerner Corporation, Kansas City, MO, United States

ARTICLE INFO

Keywords:
COVID-19
Electronic health records
Case definition

ABSTRACT

Purpose: To evaluate the number, characteristics, and outcomes of patients identified hospitalized with coronavirus disease 2019 (COVID-19) using two different case definitions.

Procedures: Electronic Health Record data were evaluated from patients hospitalized with COVID-19 through May 2020 at 52 health systems across the United States. Characteristics of inpatients with positive laboratory tests for SARS-CoV-2 were compared with those with clinical diagnosis of COVID-19 but without a confirmatory lab result.

Findings: Of 14,371 inpatients with COVID-19, 6623 (46.1 %) had a positive laboratory result, and n = 7748 (52.9 %) had only a clinical diagnosis of COVID-19. Compared with clinically diagnosed cases, those with laboratory-confirmed COVID were similar in age and sex, but differed by race, ethnicity, and insurance status. Laboratory-confirmed cases were more likely to receive certain COVID-19 therapies including hydroxychloroquine, anti-IL6 agents and antivirals ($p < 0.001$). Those with laboratory–confirmed COVID-19 had lower rates of most complications such as myocardial infarction, but higher overall mortality ($p < 0.001$).

Conclusion: We observed a two–fold difference in the number of patients hospitalized with COVID-19 depending on whether the case definition required laboratory confirmation. Variations in case definitions also led to differences in cohort characteristics, treatments, and outcomes.

1. Purpose

The coronavirus pandemic has highlighted the utility of using data from the electronic health record (EHR) to track and understand new diseases. While the EHR contains valuable information about patients with COVID-19, how to best identify COVID-19 cases using the EHR remains a challenge.

Early in the pandemic, lack of widespread availability of reliable testing, and inconsistent access to testing for the SARS-CoV-2 virus led to many COVID-19 cases being diagnosed based on clinical criteria alone. Additionally, improper specimen collection, low viral loads, and less than 100 % test sensitivity often led to false negative tests [1–3].

Prior to the pandemic, the tenth revision of International Classification of Disease (ICD-10) had only non-specific codes for coronavirus infections. In February 2020, specific ICD-10 codes were released by the World Health Organization (WHO) for COVID-19, and on April 1, the Centers for Disease Control (CDC) officially adopted the COVID-19 specific diagnosis code, U07.1 for use by US healthcare providers and coders [4,5]. Even after the release, however, the uptake and general use of this code in community practice was likely inconsistent.

EHR data have been used for a number of COVID–19 epidemiologic studies [6–8]. Understanding how different case definitions impact surveillance efforts can inform ongoing research, and may help guide = practices regarding coding of new diseases in the future. Using one of the nation's largest ongoing EHR-based COVID-19 databases, we examined the number of hospitalized patients with COVID-19 infection based on the presence of a positive laboratory test compared with the number of hospitalized patients who had only a clinical diagnosis code. We then compared characteristics of those two case groups to assess for systematic differences based on the method of case capture.

* Corresponding author at: 5323 Harry Hines Blvd., Dallas, TX 75309, United States.
E-mail address: ann.navar@utsouthwestern.edu (A.M. Navar).

https://doi.org/10.1016/j.ijmedinf.2022.104842
Received 17 May 2021; Received in revised form 16 July 2022; Accepted 4 August 2022
Available online 8 August 2022

2. Procedures

The study database was developed using de-identified EHR data from 52 geographically dispersed health systems in the US participating in Cerner Real World Data. This dataset includes all EHR data from participating health systems that opt-in to data sharing. To protect patient confidentiality, all Health Insurance Portability and Accountability Act (HIPAA) identifiers were removed; all dates were shifted up to 4 weeks (consistently forward or backward within an individual patient to preserve inter-patient temporal relationships); and the age of patients ≤17 or >89 years was reset to age 17 and 90, respectively. For this analysis, we identified one hospitalization per patient, using the most recent hospitalization for patients with multiple qualifying admissions.

This analysis included patients admitted as inpatients with COVID-19, including those in observation status. First, we identified inpatient encounters with an admission date between December 19, 2019 and June 8, 2020 (using the shifted dates after de-identification). These dates were selected to account for the date shifting in the data as noted above and therefore reflect cases from the start of the pandemic through approximately May 10, 2020. From this set of inpatient hospitalizations, encounters with a positive SARS-CoV-2 laboratory test during or within the two weeks prior to the admission were included, as were patients with one or more diagnosis codes during the hospitalization consistent with COVID-19 illness or exposure (**eTable 1**) [5]. The official SARS-Cov-2 Logical Observation Identifiers Names and Codes (LOINC) code mapping guidance from the Department of Health and Human Services [9] was used to identify laboratory codes for SARS-Cov-2 infection (**eTable 1**).

Available data included demographic information such as age, sex, race, and ethnicity. To protect patient privacy, a patient's geographic location was limited to the first digit of their ZIP code. If a patient's race was not recorded consistently in their EHR across their historical encounters, race was set to "multiple races". Encounter information for the study COVID-19 hospitalization included admission and discharge dates, vital status, medications administered, procedures, diagnoses, and laboratory data.

Within the dataset, four mutually exclusive COVID cohorts were identified based on their presumed diagnostic certainty. First, patients with at least one positive SARS-CoV-2 laboratory result during the hospital visit or in the 2 weeks prior were selected for the *Laboratory-confirmed case group*. Next, patients with at least one ICD-10 diagnosis code for COVID-19 illness were selected for the *Clinically Diagnosed case group*. These two groups formed the study's COVID-19 cases.

The remaining hospitalizations had exposure-related COVID-19 diagnosis codes but lacked both SARS-CoV-2 positive laboratory results during or 2 weeks prior to the hospital visit *and* COVID-19 clinical diagnoses during the hospital visit. These hospitalizations were further subset into two groups: (1) hospitalizations with exposure-related ICD-10 diagnosis codes during the hospital stay but no laboratory testing during or within 2 weeks of the hospitalization (i.e., *Possible cases)*, and (2) hospitalizations with exposure-related ICD-10 diagnosis codes *and* at least one negative laboratory result during the hospital stay or in the 2 weeks prior to the admission (i.e., *Probable Negative* cases).

Comorbidities at the time of admission were defined using ICD-10 diagnosis codes recorded during previous encounters in the same healthcare system in the past 3 years. Patients without any encounters in the EHR prior to hospitalization were excluded from comorbidity evaluations. Obesity was defined using a body mass index (BMI) >30 kg/m^2 or weight and height measurements taken during or prior to the hospital visit. When BMI was unavailable, ICD-10 diagnosis codes for obesity (e.g., ICD-10 Z68.36) were used. BMI was not evaluated in those <18 years of age due to lack of specific ages for these individuals. The complete list of ICD-10 diagnosis codes used to identify comorbidities is provided in **eTable 2**.

Complications of COVID-19 were identified using ICD-10 diagnosis codes, with extracorporeal membrane oxygenation (ECMO) and mechanical ventilation identified using ICD-10-PCS and Current Procedural Terminology (CPT®) procedure codes (see **eTable 3**). Medications given during the hospitalization stay were evaluated using Multum [10] medication codes (see **eTable 4**).

Descriptive statistics are presented for characteristics, treatments, complications, and outcomes for patients with clinically diagnosed COVID-19 compared with those with laboratory-confirmed disease, with t-tests used for continuous variables and Pearson chi-squared tests or Fisher's exact tests used for categorical variables when applicable.

This study was reviewed by Advarra IRB (PRO00043598).

3. Findings

3.1. Cerner COVID-19 study population

A total of 16,900 patients qualified for the COVID-19 study population across 52 centers contributing data. Within this group, 14,371 COVID-19 cases (85.0 %) had either a positive laboratory test for SARS-CoV-2 or a diagnosis of COVID-19 illness. The remaining 2529 patients (15.0 %) included n = 1488 possible cases based on the presence of a COVID-19 exposure code but without any SARS-CoV-2 laboratory test data, and n = 1041 probable negative cases who had a diagnosis for COVID-19 exposure and at least one negative SARS-CoV-2 laboratory test during the hospital visit or 2 weeks prior to admission. Characteristics of these latter two groups are presented in **eTable 5**. The regional distribution of patients in the study population is shown in **eFigure 1**. The highest concentration of patients (18.6 %) was seen in New England, followed by the South Atlantic with 15.3 % of patients, and then the Pacific region (14.0 %).

3.2. Baseline characteristics of laboratory-confirmed and clinically diagnosed COVID-19 cases

Of the overall sample of laboratory-confirmed or clinically diagnosed COVID-19 cases (n = 14,371; hereafter described as COVID-19 cases), 46.1 % (n = 6623) were identified based only on SARS-CoV-2 laboratory data whereas the remainder (n = 7748, 53.9 %) were identified based on clinical diagnoses consistent with COVID-19 infection. Among those with clinical diagnoses, n = 641 (8.3 %) had at least one negative laboratory test during the hospital stay or in the two weeks prior; the remaining 91.7 % of patients (n = 7107) had no SARS-CoV-2 laboratory testing data available within the EHR record. Conversely, of the n = 6623 patients with laboratory-confirmed disease, n = 1403 (21.2 %) had no COVID-19 diagnoses during the hospitalization. Of these, n = 865 (61.7 %) had at least one diagnosis consistent with a clinical respiratory tract illness including upper or lower respiratory disease, sepsis, or fever (see **eTable 6**). The remainder (n = 538) had other nonspecific diagnoses during their hospitalization (Fig. 1). Diagnoses for these patients that occurred in at least 10 patients are listed in **eTable 7**.

Characteristics of the 14,371 COVID-19 cases are presented in Table 1. The median age was 64 years, and 54 % were male. Just over half were white, 25.0 % were Black or African American, 3.5 % Asian or Pacific Islander, and 21.4 % were Hispanic or Latino. Among cases, n = 7303 (50.8 %) had data available from prior healthcare encounters to evaluate preexisting comorbidities. Data to evaluate obesity were available in 12,337 out of 13,642 adults 18 and over (90.4 %); n = 5566 (45.1 %) of those were obese.

3.3. Comparison of COVID-19 case groups: Differences between laboratory-confirmed and clinically diagnosed cases

3.3.1. Differences in patient characteristics

Table 1 shows characteristics of laboratory-confirmed COVID-19 cases compared with clinically diagnosed cases. Age and sex were similar between the two groups, However, variability was seen by race and ethnicity ($p < 0.001$ for both), with more Black or African American

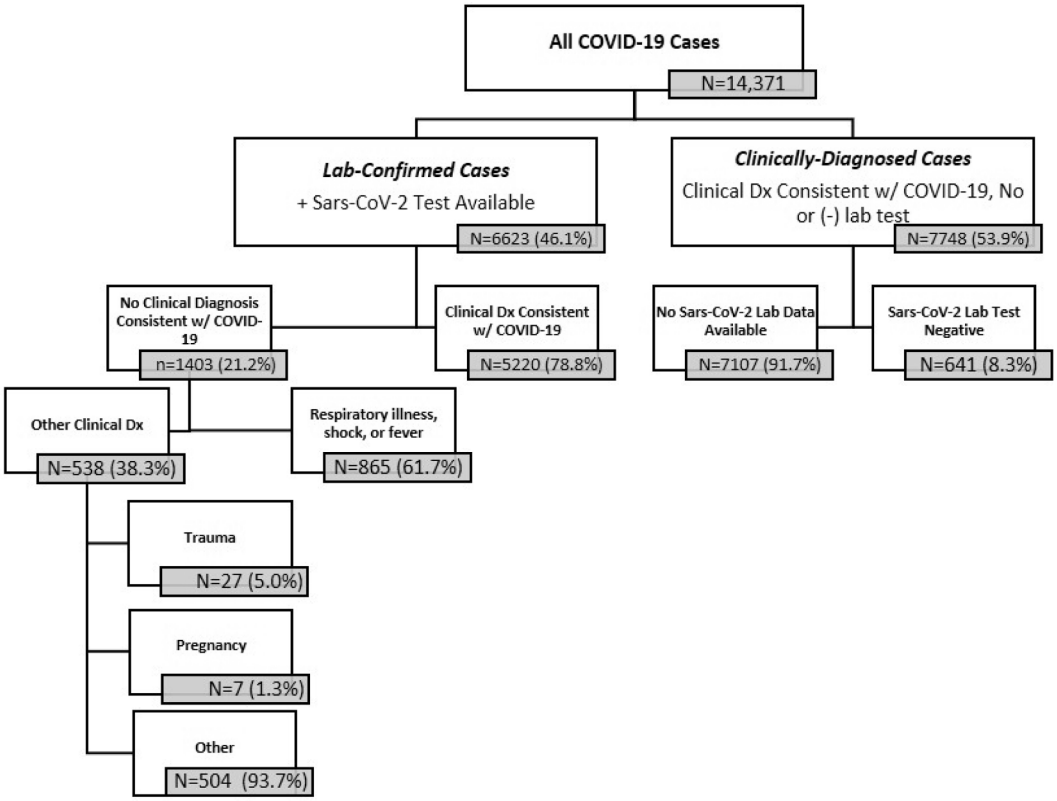


Fig. 1. Flowchart of COVID-19 Cases. This figure shows the breakdown of n = 14,371 patients hospitalized with COVID-19 infection, stratified by those with laboratory confirmation vs clinical diagnoses. Within those with laboratory confirmation, further stratification is shown by those with and without at least one diagnosis consistent with COVID-19 infection, and among those without a clinical diagnosis, those who had a diagnosis consistent with COVID-19 illness such as respiratory illness, shock, and fever.

adults in those with laboratory-confirmed disease (32.6 % vs 18.1 %) and fewer patients of Hispanic ethnicity (17.6 % vs 24.6 %). Differences between groups were also seen by insurance status. Those with laboratory confirmation had higher rates of private insurance or Medicare and less likely to have Medicaid or no insurance (p=<0.0001 for overall differences by insurance).

Among those with comorbidity data available (n = 8303), the prevalence of some comorbidities differed between those with laboratory-confirmed and clinically diagnosed infection, though differences were not consistent in any one direction. Those with laboratory-confirmed disease had higher rates of diabetes and end stage renal disease, while those with clinically diagnosed illness had higher rates of chronic respiratory diseases and coronary artery disease, among others (Table 1). There was no difference in the prevalence of heart failure, hypertension, HIV, liver disease, cancer, organ transplant status, or obesity between the two groups.

Fig. 2 demonstrates regional variability in the relative proportion of COVID-19 cases that were laboratory-confirmed vs clinically diagnosed. The proportion of cases that were laboratory-confirmed ranged from 70.1 % among states in the Western Rockies (ZIP '8') to 25.4 % in states in the Pacific region (ZIP '9'). **eFigure 2** shows how the proportion of COVID-19 cases with positive laboratory tests increased over time, from 0.44 % in the early part of the study to 55.1 % in the final week, and remained highest in Black and African American patients and lowest in white patients throughout the study period. Characteristics of patients hospitalized with a clinical diagnosis of COVID-19 also varied over time (**eTable 8)** with increases in the proportion of Black and patients and a decrease in Hispanic patients, and an increase in the proportion of patients with private insurance.

3.4. Differences in COVID-19 treatments

Treatment differences were observed in those with laboratory-confirmed disease compared with those with a clinical diagnosis (Table 1). Those with laboratory-confirmed disease were much more likely to receive chloroquine or hydroxychloroquine (63.7 % vs 28.3 %, $p < 0.0001$), an anti-IL-6 agent (9.1 % vs 2.9 %, p=<0.001), or anti-viral medication (6.1 vs 1.9 %, $p < 0.0001$) compared with those with clinical diagnoses.

3.5. Differences in clinical complications and outcomes

Rates of complications overall and by case group are shown in Table 2. Compared to those with laboratory–confirmed disease, patients with a clinical diagnosis had higher rates of venous thromboembolism, extracranial bleed, myocardial infarction, and mechanical ventilation ($p < 0.05$ for all), while acute respiratory distress syndrome was more frequently diagnosed among the laboratory-confirmed vs those with clinical diagnoses (7.0 % vs 5.4 %, $p < 0.0001$). In-hospital mortality rate was higher in the laboratory-confirmed vs clinically diagnosed cases (21.0 % vs 14.3 %, $p < 0.0001$).

4. Principal conclusions

Real-world health care databases such as EHR and administrative claims data have been critical to understanding the epidemiology of COVID-19. Unfortunately, reliable identification of patients hospitalized with COVID-19 early in the pandemic was limited by lack of availability of universal, accurate laboratory testing. In this large EHR-derived dataset of patients hospitalized with COVID-19 at over 50 health systems across the US, we found significant differences in the number and

Table 1
Characteristics and Treatment of COVID-19 Cases Overall and Stratified by the Presence or Absence of Laboratory Confirmation.

	All Cases N = 14,371	Laboratory-confirmed N = 6623	Clinically diagnosed N = 7748	p-value
Median Age, yr	64 (52–76)	64 (52–75)	64(52–76)	0.63
Sex (% male)	7742 (53.9)	3578 (54.0)	4164 (53.7)	0.64
Race				
White	7433 (51.7)	2643 (39.9)	4790 (61.8)	<0.0001
Black or African American	3555 (24.7)	2156 (32.6)	1399 (18.1)	
Asian or Pacific islander	498 (3.5)	229 (3.5)	269 (3.5)	
American Indian or Alaska Native	243 (1.7)	143 (2.2)	100 (1.3)	
Multiple race group	5 (0.03)	0 (0.0)	5 (0.1)	
Other racial group (multiple listed within the same encounter)	1732 (12.1)	980 (14.8)	752 (9.7)	
Unknown race	905 (6.3)	472 (7.1)	433 (5.6)	
Ethnicity				
Hispanic or Latino	3073 (21.4)	1166 (17.6)	1907 (24.6)	<0.0001
Not Hispanic or Latino	9257 (64.4)	4550 (68.7)	4707 (60.8)	
Unknown ethnicity	2041 (14.2)	907 (13.7)	1134 (14.6)	
Insurance				
Private/Commercial	4527 (31.5)	2320 (35.0)	2207 (28.5)	<0.0001
Medicare	4841 (33.7)	2398 (36.2)	2443 (31.5)	
Medicaid	1801 (12.5)	704 (10.6)	1097 (14.2)	
Other federal government insurance**	197 (1.4)	84 (1.3)	113 (1.5)	
Self-Pay	807 (5.6)	374 (5.6)	433 (5.6)	
Other	80 (0.6)	39 (0.6)	41 (0.5)	
Unknown insurance type	2118 (14.7)	704 (10.6)	1414 (18.2)	
Admitted from ER	12,739 (88.6)	5895 (89.0)	6844 (88.3)	<0.001
Data Available to Evaluate Baseline Comorbidities	7303 (50.8)	3273 (49.4)	4030 (52.0)	
Diabetes	3400 (46.6%)	1593 (48.7)	1807 (44.8)	0.001
Hypertension	5392 (73.8)	2453 (74.9)	2939 (72.9)	0.051
Congestive heart failure	1680 (23.0)	726 (22.2)	954 (23.7)	0.132
Coronary artery disease	2105 (28.8)	884 (27.0)	1221 (30.2)	0.002
End stage renal disease	621 (8.5)	322 (9.8)	299 (7.4)	<0.001
Asthma/Chronic bronchitis/Chronic obstructive pulmonary disease	2183 (29.9)	816 (24.9)	1367 (33.9)	<0.001
Other chronic lung disease	324 (4.4)	116 (3.5)	208 (5.2)	<0.001
Other interstitial pulmonary disease	185 (2.5)	66 (2.0)	119 (3.0)	0.011
Cancer	1175 (16.1)	501 (15.3)	674 (16.7)	0.101
HIV	101 (1.4)	41 (1.3)	60 (1.5)	0.390
Solid organ transplant	147 (2.0)	73 (2.2)	74 (1.8)	0.233
		32 (1.0)	73 (1.8)	0.003
Other immunodeficiency (besides cancer, HIV, solid organ transplant)	105 (1.4)			
Liver disease	859 (11.8)	377 (11.5)	482 (12.0)	0.560
Obesity (among those with available data, n = 12,337)	5566 (45.1)	2,865 (45.1)	2701 (45.2)	0.874
Medications Administered				
Hydroxychloroquine or Chloroquine	6414 (44.6)	4,222 (63.7)	2,192 (28.3)	<0.0001
Any Antiviral	549 (3.8)	403 (6.1)	146 (1.9)	<0.0001
Lopinavir/ritonavir	490 (3.4)	379 (5.7)	111 (1.4)	<0.001
Ritonavir	8 (0.1)	5 (0.1)	3 (0.0)	0.483
Other antiviral (not lipinavir/ritonavir or ritonavir)	88 (0.6)	47 (0.7)	41 (0.5)	0.1668
Any anti-IL6	824 (5.7)	600 (9.1)	224 (2.9)	<0.001
Tocilizumab	821 (5.7)	597 (9.0)	224 (2.9)	<0.001
Systemic steroids	4315 (30.0)	1970 (29.7)	2345 (30.3)	0.049
Vasopressor	2378 (16.5)	1347 (20.3)	1031 (13.3)	<0.001

P-value represents comparison between laboratory-confirmed vs clinically diagnosed disease. Data presented are median (25th percentile-75th percentile) for continuous variables, n (%) for categorical variables. **Includes Tricare (CHAMPUS), Department of Veterans Affairs and Other Government (Federal).

characteristics of patients hospitalized with COVID-19 depending on whether a positive laboratory test was required for case identification. The proportion of COVID-19 cases with laboratory confirmation increased over time as the SARs-CoV-2, but varied by race, geography, and insurance type/status. These differences are likely to impact epidemiologic surveillance efforts and have substantial implications for other COVID-19-related research.

Our primary finding relates to the large proportion of hospitalized COVID-19 cases that did not have laboratory confirmation available, even during the later time periods of the study when the diagnostic laboratory test for the virus was more available. Of 14,371 cases of COVID-19 infection identified, over half (53.9 %) did not have a positive SARS-CoV-2 laboratory result in the EHR. The reasons for this are likely multifactorial. First, though this database did have access to structured laboratory test results, results could be missing for patients who had laboratory tests conducted outside of the health system either prior to admission or as a "send-out" to a different laboratory during the hospital visit. Second, access to testing was at first unavailable and often limited, particularly early in the pandemic. Third, false negative tests were possible, even among patients with active disease. Importantly, though, <10 % of cases with COVID-19 that were clinically diagnosed had only negative laboratory test(s) for COVID-19. These patients could represent those with false negative tests, those with tests run to document viral clearance, or misdiagnosed cases. Without manual chart reviews, we are unable to determine the true status of these cases. However, one recently presented analysis from the FDA Sentinel System demonstrated high positive predictive value of using clinical diagnoses alone to COVID-19 patients using the EHR, suggesting that many of these cases are likely true COVID-19 infections [11].

It is possible that some of the clinically-diagnosed cases were inaccurately diagnosed. Clinically diagnosed cases were less likely to receive treatment with medications more specific for COVID-19 including

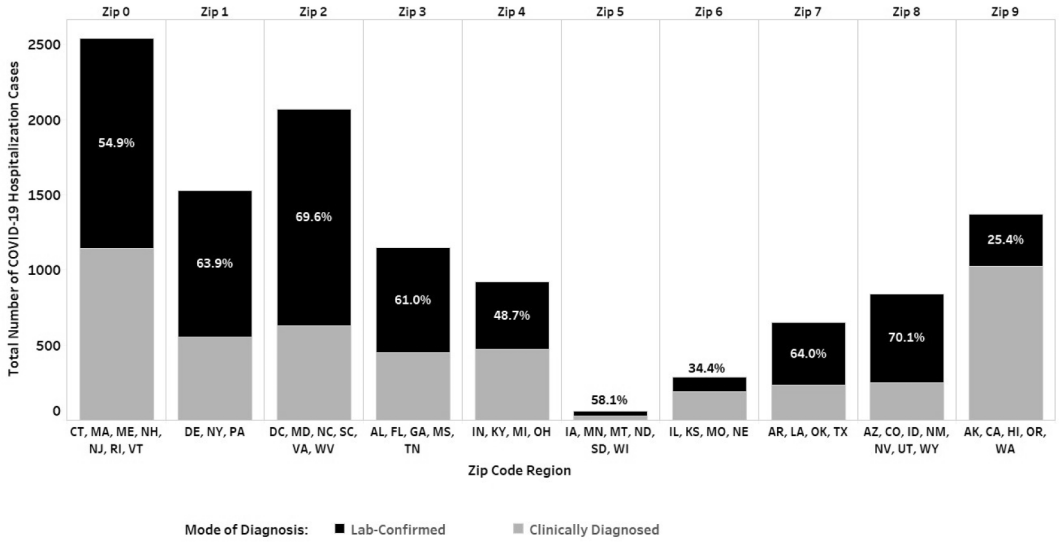


Fig. 2. COVID-19 Cases by Region, Stratified by Laboratory Confirmed and Clinically Diagnosed Cases. This figure demonstrates the geographic distribution of patients hospitalized with COVID-19 based on first digit of patient ZIP, stratified by those with a positive laboratory test compared with those who had no test or a negative test but who had a diagnosis code consistant with COVID-19 illness. Percentages represent percent of cases with at least one positive laboratory test for SARS-CoV-2.

Table 2
Complications and Outcomes of COVID-19 Cases.

Complications	All Cases N = 14,371	SARS-CoV-2 Laboratory-confirmed N = 6623	Clinically diagnosed only N = 7748	p-value
Pulmonary embolism	234 (1.6)	110 (1.7)	124 (1.6)	0.670
Venous thromboembolism	279 (1.9)	104 (1.6)	175 (2.3)	0.003
Extracranial bleed	190 (1.3)	63 (1.0)	127 (1.6)	<0.001
Myocardial infarction	642 (4.5)	221 (3.3)	421 (5.4)	<0.001
ARDS	880 (6.1)	461 (7.0)	419 (5.4)	<0.001
Respiratory failure (excluding ARDS)	5813 (40.4)	2514 (38.0)	3299 (42.6)	<0.001
Pneumonia	8407 (58.5)	3856 (58.2)	4551 (58.7)	0.531
Stroke	275 (1.9)	119 (1.8)	156 (2.0)	0.345
ECMO	11 (0.1)	1 (0.0)	10 (0.1)	0.014
Mechanical ventilation	1372 (9.5)	490 (7.4)	882 (11.4)	<0.001
Still hospitalized	2537 (17.7)	1,739 (26.3)	798 (10.3)	<0.0001
In-hospital death[+]	2021 (17.1)	1,025 (21.0)	996 (14.3)	<0.001
Median Length of stay (days) among discharged alive[++]	5.09 (3.0–8.9)	5.8 (3.2–10.0)	4.8 (2.8–8.2)	<0.0001

P-value represents comparison between laboratory-confirmed vs clinically diagnosed disease. Data presented represent n (%) for categorical variables, median (25th percentile-75th percentile) for continuous variables.

[+] Mortality rates calculated excluding those still hospitalized, denominator for this row is n = 11,834.

[++] The length of stay for the hospital visit was calculated using the inpatient hospital admission and discharge dates and times for patients who were not admitted through the emergency department. For patients who were admitted through the emergency department, length of stay is based on the emergency department admission date and time and the inpatient discharge date and time.

antivirals such as ritonavir, hydroxychloroquine, and tocilizumab, suggesting that providers may not have been as confident in the diagnosis. It is also possible that healthcare providers were influenced by the presence of a positive laboratory test and were more likely to use novel therapies. This highlights the potential importance of ensuring equitable access to testing in order to help ensure equitable access to treatments. While most patients would likely have been able to communicate a positive test from an outside lab to their provider, if testing did influence treatment, then our data also highlight the importance of ensuring complete and easy information exchange between health and laboratory systems. Most, but not all, complication rates were slightly higher in those with a clinical diagnosis. Despite this difference, the mortality rate was higher in those with laboratory diagnoses. This potentially conflicting result may be due to statistical play of chance, or may be due to misclassification of patients with less fatal alternative diagnoses.

Some laboratory-confirmed cases may have been incidentally diagnosed infections in patients who were asymptomatic from COVID-19 and hospitalized for reasons other than COVID-related illness. However, the vast majority of COVID-19 cases with laboratory confirmation also had a diagnosis code during their hospitalization consistent with either COVID-19 infection or a clinical syndrome consistent with sepsis or respiratory tract infection. Importantly, however, these data reflected the epidemiology early in the pandemic. Since that time, vaccination has dramatically lowered the case fatality rate of COVID-19, and testing is now routine for most hospitalized patients, likely increasing the relative number of incidentally detected cases. For now, CDC guidelines recommend the same ICD-10 code for those with asymptomatic incidental infection as those with active disease; future consideration should be given to developing a new code to distinguish the two [12].

Racial differences in the proportion of total COVID-19 cases with laboratory confirmation were consistently observed throughout the study period, which has the potential to impact epidemiological inferences regarding the disease state and associations with race and geography. Whether this finding is due to geographic differences in testing frequency, differential testing within a region by race, or differences in disease prevalence by race remains unknown. It is well documented that Black persons in the US were disproportionately affected by the COVID-19 pandemic, which may have contributed to higher rates of laboratory-confirmed COVID-19 in Black patients in this dataset [13,14]. However, Hispanic adults are also at higher risk of illness, and this group was not more likely in our study to have laboratory-confirmed disease [13]. Future work should evaluate whether there are systematic differences in access to testing by race or ethnic group.

EHR data can provide enormous benefits to researchers and public health officials as these data can provide a rapid, large-scale look into the current state of the pandemic and outcomes in hospitalized patients, and often include more clinical detail than administrative or claims-

based datasets. Our study highlights the inherent limitations of using the EHR to evaluate disease epidemiology, particularly for a new disease where clinical diagnostics are evolving and specific diagnostic codes are non-existent. Beyond epidemiology, differences in the accuracy of EHR-based case definitions also have the potential to impact a wide range of other domains, including assessments of healthcare utilization, clinical trial patient identification, and quality assessment and benchmarking [15].

Our study also demonstrates the critical need to address the fragmented healthcare informatics infrastructure in the US. A person receiving a laboratory test for COVID-19 at a mobile testing clinic or outside hospital may not necessarily have their results transmitted to their hospital record, and even if transferred, the result may not be entered as a structured data element accessible to EHR-based queries. Future efforts to improve medical data interoperability and increased communication between healthcare systems electronic records, will be critical to resolving this gap.

This study had several other limitations. First, only about half of patients had prior encounters in the health system where they were hospitalized to evaluate the presence of prior comorbidities. If those without prior comorbidity data also had less access to healthcare overall, or poorer control of chronic diseases, this may lead to an underestimation of patient comorbidities in a COVID-19 population and an under-appreciation of the impact of these factors on disease outcomes. Similarly, we could only query data elements captured in structured fields, preventing an analysis of clinical notes. Further, we were unable to capture experimental treatments such as remdesivir or donor plasma as these were not part of standard order lists that could be queried using structured data in this study's dataset.

Different EHR-derived case definitions result in variability in the numbers of patients considered hospitalized with COVID-19, with important differences in demographic and clinical characteristics of patients depending on the definition used. Requiring laboratory confirmation will increase the accuracy of case capture but will miss some cases. Furthermore, a COVID-19 case group identified only by laboratory confirmation may not be completely representative of the true underlying disease population due to differences in rates of laboratory testing by race, geography and other factors. On the other hand, using clinical diagnoses alone may over-estimate case numbers. Some of these challenges have likely improved with increased access to accurate and reliable testing. However, new challenges have arisen, including how to differentiate hospitalizations for COVID-19 vs incidentally detected virus in patients admitted for other reasons who undergo routine screening. Ultimately, there may not be one "best" definition of COVID-19 infection using EHR data. As researchers and regulators continue to utilize EHR- and claims-based datasets for research in the pandemic, analytic approaches that include sensitivity analyses to determine the possible impact of variation in case definition on research findings are strongly recommended.

Funding

This study was not directly funded. Indirect support for the manuscript was provided by UBC and Cerner through employee salaries for UBC and Cerner employees, respectively. Cerner provided access to the research dataset, and the analytic environment was provided by Amazon Web Services.

CRediT authorship contribution statement

Ann Marie Navar: Conceptualization, Project administration, Writing – original draft, Writing – review & editing. **Irene Cosmatos:** Conceptualization, Formal analysis, Methodology, Project administration, Writing – review & editing. **Stacey Purinton:** Data curation, Project administration, Writing – review & editing. **Janet L. Ramsey:** Writing – review & editing. **Robert J. Taylor:** Data curation, Resources, Writing – review & editing. **Rachel E. Sobel:** Formal analysis, Methodology, Resources, Writing – review & editing. **Ginger Barlow:** Formal analysis, Methodology, Writing – review & editing. **Gretchen S. Dieck:** Formal analysis, Methodology, Writing – review & editing. **Michael L. Bulgrein:** Formal analysis, Methodology, Writing – review & editing. **Eric D. Peterson:** Conceptualization, Writing – review & editing.

Declaration of Competing Interest

RJT and SP are employees of Cerner Corporation, and completed this work as part of their employment with Cerner. AMN and EDP have received consulting fees for research consulting from Cerner Corporation, outside of the scope of this work. The authors declare they have no other competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2022.104842.

References

[1] A. La Marca, M. Capuzzo, T. Paglia, L. Roli, T. Trenti, S.M. Nelson, Testing for SARS-CoV-2 (COVID-19): a systematic review and clinical guide to molecular and serological in-vitro diagnostic assays, Reprod. Biomed. Online 41 (3) (2020) 483–499.

[2] A. Scohy, A. Anantharajah, M. Bodéus, B. Kabamba-Mukadi, A. Verroken, H. Rodriguez-Villalobos, Low performance of rapid antigen detection test as frontline testing for COVID-19 diagnosis, J. Clin. Virol. 129 (2020), 104455.

[3] J. Dinnes, J.J. Deeks, A. Adriano, et al., Cochrane COVID-19 Diagnostic Test Accuracy GroupRapid, point-of-care antigen and molecular-based tests for diagnosis of SARS-CoV-2 infection, Cochrane Database Syst. Rev. 8 (2020) CD013705.

[4] Centers for Disease Control and Prevention. ICD-10-CM Official Coding Guidelines – Supplement. Coding encounters related to COVID-19 Coronavirus Outbreak. Accessed Nov 8, 2020. Available from: <https://www.cdc.gov/nchs/data/icd/interim-coding-advice-coronavirus-March-2020-final.pdf>.

[5] Centers for Disease Control and Prevention. ICD-10-CM Tabular List of Diseases and Injuries. April 1, 2020 Addenda. Accessed Nov 8, 2020. Available from: <https://www.cdc.gov/nchs/data/icd/interim-coding-advice-coronavirus-March-2020-final.pdf>. Accessed Nov 8, 2020.

[6] M.T. Oetjens, J.Z. Luo, A. Chang, et al., Electronic health record analysis identifies kidney disease as the leading risk factor for hospitalization in confirmed COVID-19 patients, PLOS One 15 (11) e0242182, doi: 10.1371/journal.pone.0242182.

[7] S. Richardson, J.S. Hirsch, M. Narasimhan, J.M. Crawford, T. McGinn, K. W. Davidson, D.P. Barnaby, L.B. Becker, J.D. Chelico, S.L. Cohen, J. Cookingham, K. Coppa, M.A. Diefenbach, A.J. Dominello, J. Duer-Hefele, L. Falzon, J. Gitlin, N. Hajizadeh, T.G. Harvin, D.A. Hirschwerk, E.J. Kim, Z.M. Kozel, L.M. Marrast, J. N. Mogavero, G.A. Osorio, M. Qiu, T.P. Zanos, Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area, JAMA 323 (20) (2020) 2052, https://doi.org/10.1001/jama.2020.6775.

[8] N. Razavian, V.J. Major, M. Sudarshan, J. Burk-Rafel, P. Stella, H. Randhawa, S. Bilaloglu, J. Chen, V. Nguy, W. Wang, H. Zhang, I. Reinstein, D. Kudlowitz, C. Zenger, M. Cao, R. Zhang, S. Dogra, K.B. Harish, B. Bosworth, F. Francois, L. I. Horwitz, R. Ranganath, J. Austrian, Y. Aphinyanaphongs, A validated, real-time prediction model for favorable outcomes in hospitalized COVID-19 patients, npj Digital Med. 3 (1) (2020).

[9] LOINC In Vitro Diagnostic (LIVD) Test Code Mapping for SARS-CoV-2 Tests. Available from: <https://www.cdc.gov/csels/dls/sars-cov-2-livd-codes.html>. Accessed Nov 8, 2020.

[10] Multum [Internet]. Denver (CO): Cerner Corporation. Solutions; [cited 2018 March 21]. Available from: <https://www.cerner.com/solutions/drug-database>. Accessed Nov 8, 2020.

[11] S. Kluberg, Validation of claims-based algorithms to identify hospitalized COVID-19 events within the FDA Sentinel System, Research findings presented at the meeting of the International Conference for Pharmacoepidemiology (ICPE) Special COVID-19 Sessions, Virtual event. December 3, 2020.

[12] ICD-10-CM Official Coding and Reporting Guidelines April 1, 2020 through September 30, 2020. Available from: <https://www.cdc.gov/nchs/data/icd/COVID-19-guidelines-final.pdf>. Accessed Nov 8, 2020.

[13] S. Sze, D. Pan, C.R. Nevill, et al., Ethnicity and clinical outcomes in COVID-19: a systematic review and meta-analysis, EClinicalMedicine 2930 (2020), 100630.
[14] Centers for Disease control. COVID-19 Hospitalization and Death by Race/Ethnicity. Available from: <https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-race-ethnicity.html>. Accessed Nov 8, 2020.
[15] R.A. Verheij, V. Curcin, B.C. Delaney, M.M. McGilchrist, Possible sources of bias in primary care electronic health record data use and reuse, J. Med. Internet Res. 20 (5) (2018), e185.