Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Development and Validation of Algorithms to Identify COVID-19 Patients Using a US Electronic Health Records Database: A Retrospective Cohort Study

Carolyn A Brown[1], Ajit A Londhe[1], Fang He[1], Alvan Cheng[1], Junjie Ma[1], Jie Zhang[1], Corinne G Brooks[1], J Michael Sprafka[1,2], Kimberly A Roehl[1], Katherine B Carlson[1,3], John H Page[1]

[1]Center for Observational Research, Amgen, Inc., Thousand Oaks, CA, USA; [2]Woodford Research Associates, Thousand Oaks, CA, USA; [3]Now with R&D Strategy, Moderna Inc., Cambridge, MA, USA

Correspondence: Carolyn A Brown; John H Page, Center for Observational Research, Amgen, Inc., 1 Amgen Center Drive, B38-4B, Thousand Oaks, CA, 91320, USA, Tel +1-818-482-9477; +1-805-490-5527, Email cbrown14@amgen.com; Jopage@amgen.com

Introduction: In order to identify and evaluate candidate algorithms to detect COVID-19 cases in an electronic health record (EHR) database, this study examined and compared the utilization of acute respiratory disease codes from February to August 2020 versus the corresponding time period in the 3 years preceding.

Methods: De-identified EHR data were used to identify codes of interest for candidate algorithms to identify COVID-19 patients. The number and proportion of patients who received a SARS-CoV-2 reverse transcriptase polymerase chain reaction (RT-PCR) within ±10 days of the occurrence of the diagnosis code and patients who tested positive among those with a test result were calculated, resulting in 11 candidate algorithms. Sensitivity, specificity, and likelihood ratios assessed the candidate algorithms by clinical setting and time period. We adjusted for potential verification bias by weighting by the reciprocal of the estimated probability of verification.

Results: From January to March 2020, the most commonly used diagnosis codes related to COVID-19 diagnosis were R06 (dyspnea) and R05 (cough). On or after April 1, 2020, the code with highest sensitivity for COVID-19, U07.1, had near perfect adjusted sensitivity (1.00 [95% CI 1.00, 1.00]) but low adjusted specificity (0.32 [95% CI 0.31, 0.33]) in hospitalized patients.

Discussion: Algorithms based on the U07.1 code had high sensitivity among hospitalized patients, but low specificity, especially after April 2020. None of the combinations of ICD-10-CM codes assessed performed with a satisfactory combination of high sensitivity and high specificity when using the SARS-CoV-2 RT-PCR as the reference standard.

Keywords: COVID-19, SARS-CoV-2, epidemiology, verification bias, validation

Introduction

In late 2019, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a novel coronavirus of zoonotic origin, was identified in Wuhan, China.[1] The virus and the disease it causes, coronavirus disease 2019 (COVID-19), have since spread across the globe, infecting tens of millions of individuals and resulting in over four million deaths.[2] In the United States (US), the first official COVID-19 case was confirmed by the Center for Disease Control and Prevention (CDC) on January 20, 2020.[3] Although cases without travel history to China did not emerge until late February, data suggest that there was sustained community transmission beginning before February 2020.[4] COVID-19 is theorized to be composed of an early viral replication phase followed by an immune dysregulation phase.[5] It is during the latter phase that many patients are hospitalized.

Real-world data (RWD), such as data from electronic health records (EHR) and insurance claims, can play an important role in COVID-19 research to better understand the disease on a population level and evaluate potential interventions. RWD typically uses coded information, such as diagnosis codes, laboratory findings, and documentation of

Received: 29 January 2022
Accepted: 25 April 2022
Published: 23 May 2022

prescription medication use, to identify and track COVID-19 patients; however, the pandemic was present in the US before the availability of any specific diagnostic codes for COVID-19, making identification of cases early in the course of the pandemic challenging. We designed this study with three primary objectives: 1) examine and compare the utilization of acute respiratory disease codes from February to August 2020 versus the corresponding time period in the 3 years immediately preceding 2020; 2) design algorithms to identify COVID-19 in the outpatient and inpatient settings on the basis of the preceding information; and 3) estimate the operating characteristics (including sensitivity and specificity) of these algorithms by clinical setting (outpatient versus inpatient) and time period (January to March 31, 2020 versus April 1, 2020 to August 30, 2020).

Methods

Data Source and Study Population

The study was conducted using the Optum® COVID-19 EHR database, which is a subset of the pan-therapeutic Optum® EHR database (PanTher). The Optum® PanTher EHRs are derived from 53 integrated delivery networks from diverse geographies in the US, including more than 700 hospitals and 7000 clinics across the US. The data are certified as de-identified following HIPAA (Health Insurance Portability and Accountability Act) statistical de-identification rules and managed according to Optum customer data use agreements. Clinical and administrative data are obtained from both inpatient and ambulatory EHRs, practice management systems and numerous other internal systems; and are processed, normalized, and standardized across acute inpatient stays and outpatient visits. Data elements include, but are not limited to, patient demographic information, medications prescribed and administered, laboratory results, and coded diagnoses and procedures. The Optum® COVID-19 EHR database was constructed to include patients with a COVID-19-specific and/or -related diagnosis code (screening, exposure, systemic and respiratory symptom/syndrome, or coded diagnosis of COVID-19 or other coronavirus infection) and patients who received a SARS-CoV-2 laboratory test. A complete list of the diagnosis codes and laboratory tests required for inclusion in the database is presented in Supplementary Table 1.

Identification of Codes of Interest

To inform codes that are used in real-world clinical practice for COVID-19, the study described the use of various COVID-19-specific and -related diagnosis, symptom and screening codes over the time period from February 1, 2020 to August 30, 2020 (Table 1). Codes of interest included diagnosis codes of COVID-19 and other coronavirus infections, codes for respiratory symptoms and syndromes, and COVID-19 exposure and screening codes. In addition, the utilization of these same codes from 2017 to 2019 in the Optum PanTher database was characterized and compared to the 2020 data to provide expected background rates.

To help select candidate codes to identify COVID-19 patients, for each diagnosis code, we calculated the following statistics: 1) the number and proportion of patients who received a SARS-CoV-2 reverse transcriptase polymerase chain reaction (RT-PCR) within ± 10 days of the occurrence of the diagnosis code and 2) the number and proportion of patients who tested positive among those who had their test result documented. The analyses were conducted in the overall study population as well as subgroups stratified by age at diagnosis (0–4, 5–11, 12–17, and 18+), by care setting at diagnosis (inpatient, outpatient), and by calendar time period to help inform the need for separate algorithms in specific sub-populations.

Design of Candidate Algorithms to Identify COVID-19

Based on the sensitivities, specificities, and positive predictive values of individual COVID-19-specific and -related diagnosis and symptom codes, those with high values were logically combined into different algorithms. We also made accommodations for the late availability of the COVID-19-specific ICD-10-CM code, U07.1, and the frequent use of coronavirus codes in the pediatric population. Any patients without a clinical symptom, clinical code for COVID-19 or documented SARS-CoV-2 viral test result were excluded from the study.

Table 1 Description of Algorithms for COVID-19 Assessed

ICD-10-CM Code	**ICD-10-CM Definition**	
U07.1	COVID-19	
J12.81	Pneumonia due to SARS-associated coronavirus	
J12.89	Other viral pneumonia	
J20.8	Acute bronchitis due to other specified organisms	
J22	Unspecified acute lower respiratory infection	
J40	Bronchitis, not specified as acute or chronic	
J80	Acute respiratory distress syndrome	
J98.8	Other specified respiratory disorders	
B34.2	Coronavirus infection, unspecified	
B97.29	Other coronavirus as the cause of diseases classified elsewhere	
Algorithm	**Descriptive Name**	**Definition**
Algorithm 1*	COVID-19 Diagnosis	U07.1 (first created February 20, 2020)
Algorithm 2*	SARS-associated pneumonia	J12.81
Algorithm 3*	Any coronavirus	U07.1 or B97.29 or B34.2
Algorithm 4	COVID-19 and pre-May coronavirus	U07.1 and (B97.29 *or* B34.2 before May 2020)
Algorithm 5*	Any coronavirus plus respiratory condition	(U07.1 or B97.29 or B34.2) and (J12.89 or J20.8 or J22 or J40 or J80 or J98.8) during the same encounter
Algorithm 6	ARDS or other viral pneumonia	J80 or J12.89
Algorithm 7	COVID-19 or pre-May any coronavirus plus respiratory condition or SARS-associated pneumonia	U07.1 anytime or (J12.81 or [(B97.29 or B34.2) and (J12.89 or J20.8 or J22 or J40 or J80 or J98.8)] between Feb 1 and Apr 30)
Algorithm 8	COVID-19 or pre-May SARS-associated pneumonia or pre-May any coronavirus	U07.1 anytime or (J12.81 or B97.29 or B34.2 between Feb 1 and Apr 30)
Algorithm 9	COVID-19 or pre-May ARDS, viral/SARS-associated pneumonia, or any coronavirus	U07.1 anytime or (J80 or J12.81 or J12.89 or B97.29 or B34.2 between Feb 1 and Apr 30)
Algorithm 10*	COVID-19, ARDS, or viral/SARS-associated pneumonia, or pre-May any coronavirus	U07.1 anytime or (J80 or J12.81 or J12.89 anytime) or (B97.29 or B34.2 between Feb 1 and Apr 30)
Algorithm 11*	ARDS or viral/SARS-associated pneumonia	J80 or J12.81 or J12.89

Notes: *Indicates Algorithm is described in main text of manuscript; all other Algorithms are described in the Supplementary Material.
Abbreviations: SARS, severe acute respiratory syndrome; ARDS, acute respiratory distress syndrome.

A total of 11 algorithms were defined and are summarized in Table 1. Algorithm 1 is defined by use of U07.1 (COVID-19) anytime during the study period. Algorithm 2 is defined by the use of codes for pneumonia due to SARS-associated coronavirus. Algorithm 3 is defined by use of codes for at least one coronavirus code (U07.1, B97.29, or B34.2) at any time during the study period. Algorithm 4 is defined by use of U07.1 at any time and one of the other coronavirus codes (B97.29, or B34.2) before May 2020. Algorithm 5 is defined by use of codes for algorithm 3 used in combination with one of various respiratory condition codes (J12.89, J20.8, J22, J40, J80, or J98.8). Algorithm 6 is defined by use of codes for acute respiratory distress syndrome (J80) or viral pneumonia (J12.89). Algorithms 7–10 recognized the lack of availability of a specific COVID-19 code prior to April, and are defined by use of the U07.1 code at any time during the study period or use of various alternative codes before May of 2020 (Table 1). Algorithm 11 is defined by use of codes for acute respiratory distress syndrome (J80) or viral pneumonia or SARS pneumonia (J12.81 or J12.89).

Validation of Candidate Algorithms

Candidate algorithms were assembled based on different combinations of codes of interest identified in the aforementioned step. The index date for this analysis was the earliest date in the study period a person qualified for one of the candidate algorithms. For patients with multiple dates satisfying different algorithm criteria, the earliest was used. To evaluate the combined performance of each algorithm, we estimated the sensitivity, specificity, positive likelihood ratio (PLR), negative likelihood ratio (NLR), positive predictive value (PPV) and negative predictive value (NPV) in a subgroup of patients for whom confirmed results from SARS-CoV-2 RT-PCR were available, using the test results for SARS-CoV-2 RT-PCR as the reference standard. Among patients who met a candidate algorithm, test results that occurred on or within 14 days of meeting the algorithm definition were used to classify the case as a true or false positive. To facilitate the estimation of confidence intervals and the weighting (see next step), we estimated sensitivity, specificity, likelihood ratios, and predictive values using log-linear regression models. When estimates of upper confidence limits for sensitivity, specificity, or predictive values exceeded 1, we constrained the upper limit at 1.

Adjustment for Verification Bias

Not all patients with a suspicion of COVID-19 had verification of infection by a SARS-CoV-2 RT-PCR. The probability of verification was likely influenced by patient, practice, and period influences. We therefore adjusted for potential verification bias by a weighted analysis; weights were calculated as the reciprocal of the estimated probability of verification in each verified study participant. We estimated the weights using predicted probabilities generated after estimating a logistic regression model for verification by either SARS-CoV-2 RT-PCR or antigen testing among the entire study sample. Covariates included were month of index date, age at diagnosis, sex, census division, integrated delivery network (IDN) inclusion, ventilation status, presence of baseline characteristics (receipt of chemotherapy, chronic obstructive pulmonary disease (COPD), heart failure, sickle cell disease, obesity), and evidence of individual clinical symptoms/findings associated with COVID-19 (pharyngitis, upper respiratory tract infection, headache, fever, dyspnea, malaise, loss of taste or smell, and chest x-ray findings). Weights generated from the logistic regression were applied to the study population; and sensitivity, specificity, likelihood ratios, and predictive values were re-estimated. We estimated 95% bias-corrected and accelerated bootstrap (BCa) confidence intervals from respective 1000 bootstrap estimates (stratified by study period and inpatient status).[6] If the confidence interval could not be estimated due to small sample size within a category, the confidence interval was left blank. Due to differences in testing availability in the early months of the pandemic as well as differences between hospitalized patients and outpatients, the analysis was performed in four distinct groups: 1) hospitalized patients before April 2020; 2) outpatients before April 2020; 3) hospitalized patients on or after April 1, 2020; and 4) outpatients on or after April 1, 2020. Findings from groups 1 and 2 (ie, patients with an Index Date before April 2020) are described in the Supplementary Tables 2 and 3. The six algorithms described in the main text of the manuscript are less time constrained and more likely to be useable for future research examining patients with symptoms suggestive of COVID-19, however sensitivity, specificity, likelihood ratios, and predictive values for all the algorithms are described in Supplementary Tables 2 and 3, Table 1.

Other Analysis

To describe the study population, descriptive statistics on demographic and baseline clinical characteristics were calculated for all patients in the COVID-19 database.

Results

Usage of Diagnosis Codes by Time Period

Patient characteristics are described by clinical setting of presentation in Table 2. Early in the pandemic (January to March 2020), the most commonly used diagnosis codes related to COVID-19 diagnosis were R06 (Dyspnea) and R05

Table 2 Descriptive Characteristics of Cohort

	Before April 2020		On or After April 1, 2020	
	Inpatients, n=5646	Outpatients, n=5119	Inpatients, n=26,158	Outpatients, n=97,012
Demographics				
Age at index (median, 10th–90th percentile)	60 (32–79)	49 (23–72)	60 (31–80)	46 (21–71)
Female N, (%)	2443 (43.3)	2840 (55.5)	12,713 (48.6)	56,741 (58.5)
Census Track N, (%)				
East North Central	1575 (27.9)	1230 (24.0)	6444 (24.6)	24,739 (25.5)
East South Central	164 (2.9)	51 (1.0)	2168 (8.3)	2464 (2.5)
Middle Atlantic	2288 (40.5)	1699 (33.2)	6597 (25.2)	20,653 (21.3)
Mountain	3 (0.1)	489 (9.6)	23 (0.1)	4448 (4.6)
New England	272 (4.8)	416 (8.1)	1738 (6.6)	7422 (7.7)
Pacific	244 (4.3)	195 (3.8)	1002 (3.8)	3845 (4.0)
South Atlantic/West South Central	432 (7.7)	444 (8.7)	4507 (17.2)	14,490 (14.9)
West North Central	530 (9.4)	444 (8.7)	2891 (11.1)	15,539 (16.0)
Other/Unknown	138 (2.4)	151 (2.9)	788 (3.0)	3412 (3.5)
Baseline Characteristics N (%)				
Obese	2921 (51.7)	2281 (44.6)	13,647 (52.2)	43,036 (44.4)
Diabetes	2124 (37.6)	914 (17.9)	10,042 (38.4)	14,859 (15.3)
COPD	1050 (18.6)	455 (8.9)	4189 (16.0)	4743 (4.9)
Heart failure	1041 (18.4)	362 (7.1)	4394 (16.8)	3381 (3.5)
Sickle cell disease	21 (0.4)	5 (0.1)	63 (0.2)	58 (0.1)
On chemotherapy	139 (2.5)	67 (1.3)	430 (1.6)	810 (0.8)
COVID-19 Testing, N, (%)				
Received viral RNA test result	4070 (72.1)	2486 (48.6)	22,873 (87.4)	59,141 (61.0)
Positive viral RNA test*	3540 (62.7)	1992 (38.9)	18,626 (71.2)	43,076 (44.4)
Received antigen test result	36 (0.6)	84 (1.6)	172 (0.7)	1166 (1.2)
Positive antigen test**	3 (8.3)	4 (4.8)	22 (12.8)	206 (17.7)
Qualified for Algorithm, N, (%)				
Algorithm 1	3087 (54.7)	1590 (31.1)	24,610 (94.1)	94,935 (97.9)
Algorithm 2	169 (3.0)	32 (0.6)	501 (1.9)	228 (0.2)
Algorithm 3	4538 (80.4)	4380 (85.6)	24,698 (94.4)	95,930 (98.9)
Algorithm 4	4536 (80.3)	4376 (85.5)	24,664 (94.3)	95,462 (98.4)
Algorithm 5	3903 (69.1)	2691 (52.6)	17,111 (65.4)	17,300 (17.8)
Algorithm 6	4487 (79.5)	1323 (25.8)	17,547 (67.1)	8631 (8.9)
Algorithm 7	4098 (72.6)	3081 (60.2)	24,642 (94.2)	94,991 (97.9)
Algorithm 8	4541 (80.4)	4384 (85.6)	24,670 (94.3)	95,470 (98.4)
Algorithm 9	5483 (97.1)	4926 (96.2)	25,031 (95.7)	95,788 (98.7)
Algorithm 10	5646 (100.0)	5119 (100.0)	26,129 (99.9)	96,560 (99.5)
Algorithm 11	4674 (82.8)	1525 (29.8)	17,613 (67.3)	8702 (9.0)

Notes: * Proportion of those testing positive among those who received a viral RNA test result; **Proportion of those testing positive among those who received an antigen test result.

(Cough). In the latter half of March and early April, the use of code U07.1 (Emergency use of U07.1 | COVID-19, created February 20, 2020) increased substantially and usage remained high until May. As the pandemic progressed, the use of the R05 and U07.1 codes declined, and diagnosis codes Z20.828 (Contact with and [suspected] exposure to other viral communicable diseases) and Z11.59 (Encounter for screening for other viral diseases) became the most frequently utilized until the end of August (Figure 1). Throughout the study period, of the codes examined in this study, the least commonly used diagnosis codes among confirmed COVID-19 patients were J12.81 (Pneumonia due to SARS associated coronavirus), B34.2 (Coronavirus infection, unspecified), and B97.29 (other coronavirus infection, classified elsewhere).

Compared to the utilization of these codes in 2017–2019, usage of J20.8 (acute bronchitis) and J40 (Bronchitis, unspecified) were much lower during the same period in 2020. Higher usage of B97.29 and J22 (Acute lower respiratory tract infection) was observed in 2020 starting in March, but the usage of these codes decreased afterwards to levels similar to those observed in 2017–2019. Increased usage of diagnosis codes J12.89 (Viral pneumonia) and J80 (Acute respiratory distress syndrome) were also observed in March 2020 and remained higher compared to their usage in 2017–2019 through the same period (Figure 2).

COVID-19 Diagnosis Code Usage with Test Positivity Rate

The proportion of all patients with a positive test among those with SARS-CoV-2 test results ranged between 4.2% and 93.9% for evaluated diagnosis codes, with the highest among those with a diagnosis code of J12.81 (Pneumonia due to SARS-associated coronavirus) and lowest among patients with diagnosis code Z11.59 (Encounter for screening for other viral diseases). This remained consistent for patients aged 18+; for patients aged 17 and under, this proportion was highest among those with code U07.1. By care setting, inpatients with diagnosis code J12.81 had the highest proportion of positive test among those with test results, while outpatients with diagnosis code J12.89 (other viral pneumonia) had the highest proportion. Receipt and positivity findings from SARS-CoV-2 RNA tests by diagnosis code, age, and month are shown in Supplementary Figures 1 and 2.

Diagnostic Statistics of Various Algorithms

Diagnostic statistics differed by algorithm, by time period assessed (before April 2020 versus after), and by clinical setting (inpatients versus outpatients), see Figure 3 and Supplementary Material. On or after April 1, 2020 the U07.1 code (COVID-19) had high adjusted sensitivity (inpatient=1.00 [95% CI 1.00, 1.00]; outpatient=1.00 [95% CI 1.00,

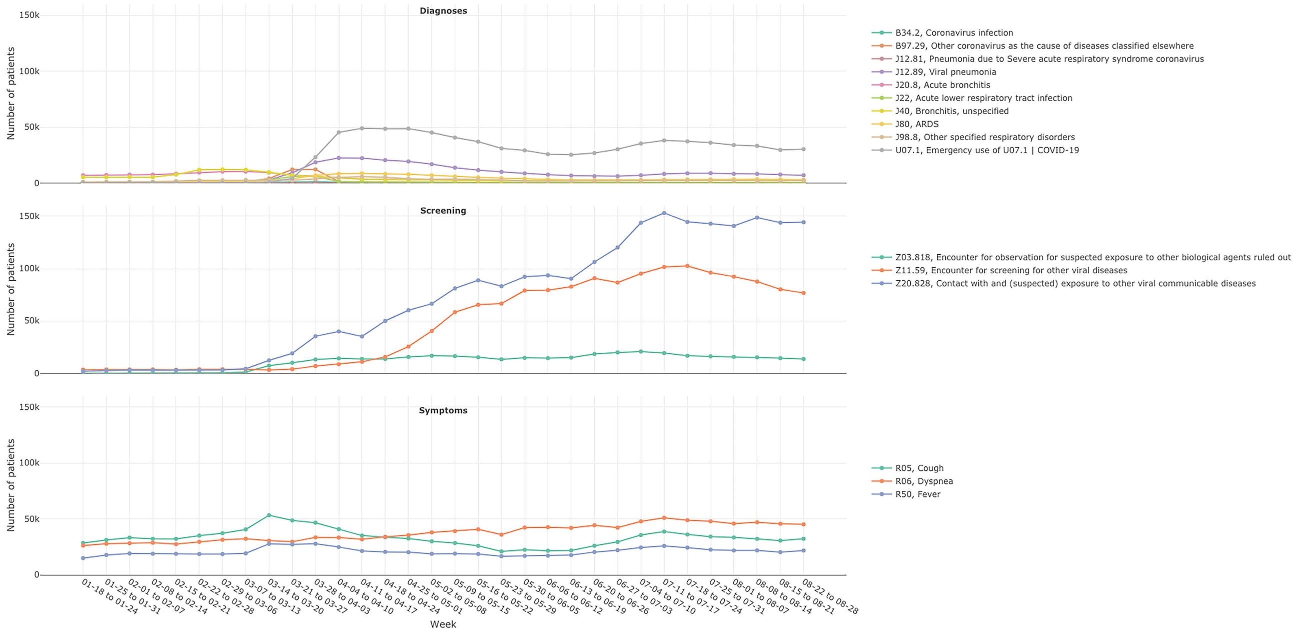

Figure 1 Time series of COVID-19 related diagnosis codes in 2020.

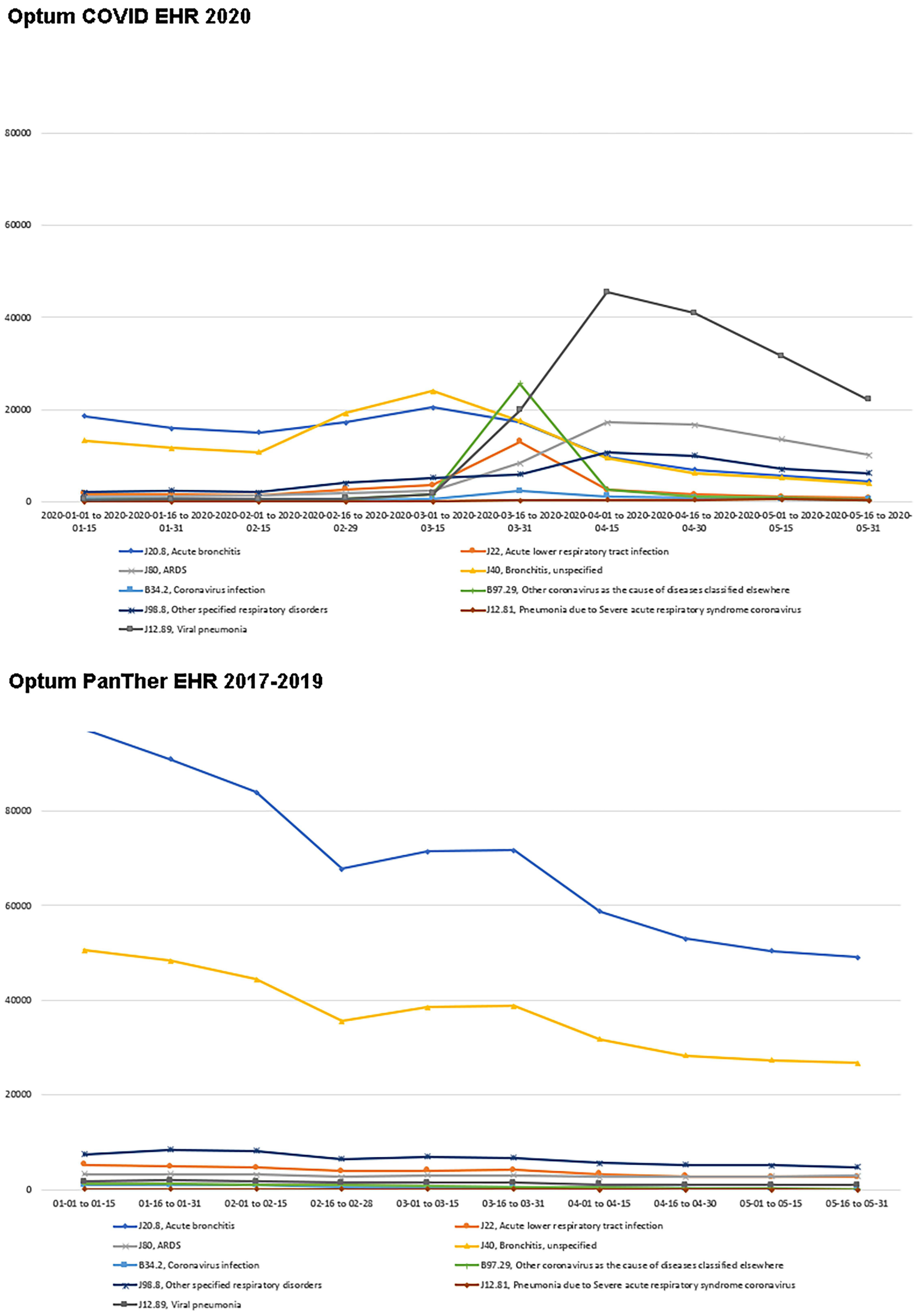

Figure 2 COVID-19-related codes in Optum COVID EHR in 2020 and Optum PanTher EHR 2017–2019.

1.00]) but low adjusted specificity (inpatient=0.32 [95% CI 0.31, 0.33]; outpatient=0.05 [95% CI 0.05, 0.06]) in both hospitalized and non-hospitalized patients. SARS-associated pneumonia demonstrated poor sensitivity and near perfect specificity across all patient groups and time frames. ARDS or viral/SARS-associated pneumonia showed higher sensitivity and lower specificity among inpatients (adjusted sensitivity=0.70 [95% CI 0.69, 0.71]; adjusted specificity=0.48 [95% CI 0.46, 0.49]) compared to outpatients (adjusted sensitivity= 0.11 [0.10, 0.11]; adjusted specificity=0.95 [95% CI 0.95, 0.95]). The positive likelihood ratios (PLRs) were close to 1 for most algorithms, with the exception of the

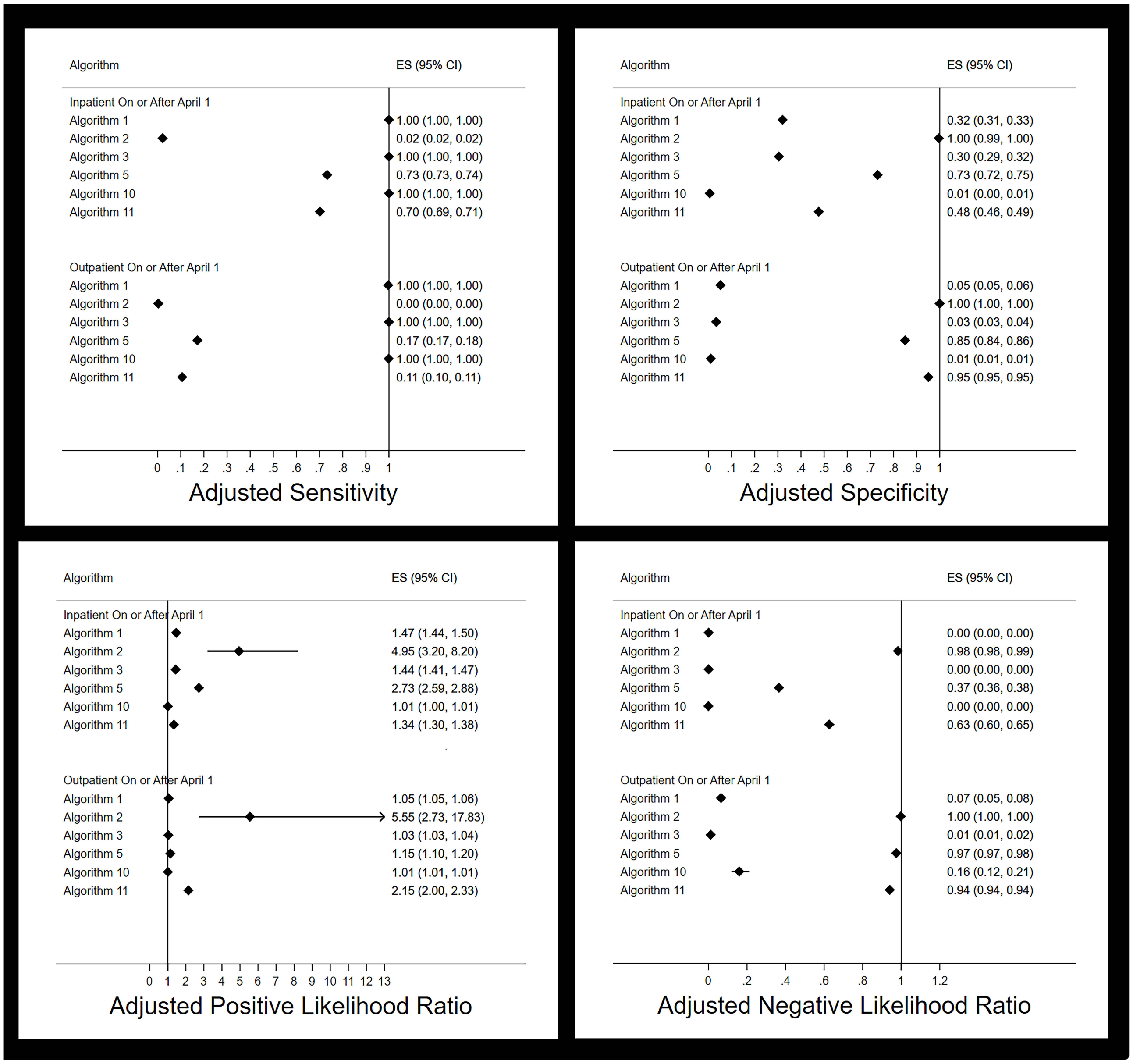

Figure 3 Adjusted performance of selected algorithms to detect COVID-19.

algorithms based on SARS-associated pneumonia, any coronavirus plus respiratory condition among inpatients, and ARDS or viral/SARS-associated pneumonia among outpatients. The adjusted results are presented in Figure 3 and the unadjusted are shown in the Supplementary Material.

Discussion

These data demonstrate that during the first eight months of 2020, there was increased use of ICD-10-CM codes for "other coronavirus", "other viral pneumonia", acute respiratory distress syndrome, and cough compared to corresponding months of 2017–2019. Health care institutions began using the specific code for COVID-19 (U07.1) in March 2020 after its creation in February 2020.[7] We found that some of these codes as well as others including "Pneumonia due to SARS-associated coronavirus" had high positive predictive value for confirmed infection with SARS-CoV-2 and were used as the basis for candidate algorithms to identify patients with COVID-19.

Using SARS-CoV-2 RT-PCR as the reference standard, we found that algorithms based on the code for COVID-19, U07.1 (Algorithm 1), had high sensitivity among hospitalized patients, but low specificity, especially after April 2020. The ICD-10-CM code for pneumonia due to SARS-associated coronavirus, J12.81 (Algorithm 2), had very high specificity, but given the low prevalence among patients with COVID-19, low sensitivity. Among the algorithms evaluated, we did not find any with both high sensitivity and high specificity, when using SARS-CoV-2 RT-PCR as the reference standard. The only evaluated algorithm with high positive likelihood ratio was Algorithm 2, the code for pneumonia due to SARS-associated coronavirus (after April 2020). A previous study reported by Kadri et al evaluated the performance of the ICD-10-CM code for COVID-19 (U07.1), for selected hospitals in a US healthcare claims database, for inpatient discharges between April 1 and May 31, 2020, and found that this code had a sensitivity of 0.98 and specificity of 0.99.[8] This was contrasted by a study by Bhatt et al among hospitalized patients at a single tertiary care facility that found that U07.1 had a sensitivity of 49.2% and a specificity of 99.4%.[9] The current study, based on electronic health records did not confirm the performance of the U07.1 code found in either study. There is a non-zero probability that clinicians used the U07.1 code to indicate the possibility of COVID-19, instead of confirmed diagnosis. This is especially likely among patients outside the hospital setting, where many patients did not appear to have interacted with a physician. Crabb et al evaluated the performance of ICD-10-CM codes for selected symptoms (fever, cough, and dyspnea) among COVID-19 patients in both inpatient and outpatient settings in a single large medical institution, and found sensitivities of 0.24 to 0.44, and specificities of 0.88 to 0.98, more in keeping with findings in this study.[10]

ICD coding is an important part of healthcare operations. During the COVID-19 pandemic, it was used to track cases, evaluate care patterns, and assess health outcomes. Based on the results of this study, ICD coding for COVID-19 lacks sufficient accuracy in US nation-wide electronic health record data. Based on the imperfect specificity of the algorithms evaluated, and assuming study designs with non-differential capture of ICD-10 coding for COVID-19, future EHR-based studies evaluating therapies or exposures as predictors of COVID-19 will likely have relative risk or risk difference measures that are biased towards the null.[11] Also, assuming non-differential measurement errors, future studies evaluating COVID-19 as a predictor of disease outcomes, in EHR systems, will likely suffer biases towards the null.[12]

The strengths of this research include the study size, and geographical coverage. These data represent a wide cross-section of patients with COVID-19 over the initial 9–10 months of the COVID-19 pandemic across the United States, with a diverse mix of insurance types, socioeconomic status, and demographic factors. To our knowledge, this is first US study to evaluate the performance of multiple algorithms to identify patients with COVID-19 using electronic health records. This is also one of the first studies of real-world data to adjust for verification bias. When the probability of verification with the reference standard is related to information about the true disease status, verification bias is present and can distort the naïve performance characteristics of the algorithm(s). Although the adjustment for verification bias did not change interpretation of the estimates of algorithm performance in this study, we believe that adjusting for verification bias is a practice to be encouraged in other studies of algorithm performance when using electronic health records.

Although real-time RT-PCR along with other nucleic acid amplification tests (NAATs) are currently considered as an approximate reference standard by the CDC, these tests are not perfect. Performance is dependent on sampling technique and specimen type,[13–15] and both clinician- and patient-related factors are associated with false-negative findings.[16] Further, some investigators were able to detect SARS-CoV-2 RNA on bronchoalveolar lavage in up to 16% of the patients negative with nasopharyngeal testing.[17] It should also be noted that patients are more likely to have virus shedding in the early symptomatic phase than in the later immune dysregulation phase when patients are more likely to hospitalized.[5,18] Although real-time RT-PCR may remain positive after virus shedding has ceased, a positive correlation exists between the probability of viral shedding and the likelihood of testing positive with SARS-CoV-2 RT-PCR.[18] The study by van Kampen et al also revealed that depending on the timing of testing in relation to onset of symptoms, patients hospitalized with COVID-19 had a clear non-zero probability of testing negative on the SARS-CoV-2 RT-PCR test when using the conventionally used cycle thresholds.[18] False positives also occur.[19,20] Depending on the nature of the errors in the standard, the use of an imperfect reference standard may result in the algorithms appearing either better or worse on sensitivity as well as specificity than they really are.[21]

Although we adjusted for verification bias, we likely did not capture all predictors of verification and we are therefore not able to guarantee the absence of residual verification bias. Further, the composition of patients in our study sample is unlikely to be generalizable to the full spectrum of patients presenting to clinicians with respiratory symptoms consistent with COVID-19. Over 90% of the patients in the study sample had an ICD-10-CM code for COVID-19, U07.1 (Table 2). This implies that either that many patients were inappropriately coded (overuse of the U07.1 code), or the study population was somehow enriched with patients with COVID-19-like presentations.

Conclusions

In summary, we did not find any combination of ICD-10-CM codes that performed with a satisfactory combination of high sensitivity and high specificity, when using the results of SARS-CoV-2 RT-PCR as the reference standard, in a subset of a US electronic medical record system. Additional research is needed to determine whether the results of SARS-CoV-2 RT-PCR can be relied on to be used as a reference standard for the diagnosis of COVID-19 or whether the U07.1 code is being misused, especially in outpatient settings. It will also be beneficial to replicate this research in other clinical settings. Clinicians and public health institutions should be cautious of relying on ICD-10-CM codes to identify patients with COVID-19 in US electronic health record systems.

Data Sharing Statement

The data were used under license for this study with restrictions that do not allow for the data to be redistributed or made publicly available. No additional data are available.

Ethics Approval

Independent IRB Advarra Inc. determined this study is exempt from ethics review under the 45 CFR 46.104(d) (4) policy of the USA Department of Health and Human Services.

Author Contributions

JHP, AAL, CGB, JZ, CAB, AC, FH, JM, KAR, KBC, JMS: study concept & design. JHP, CGB, AAL, FH, KAR: acquisition of data. JHP, AAL, CGB, JZ, CAB, AC, FH, JM, KAR, KBC, JMS: interpretation of data & drafting of manuscript. All authors contributed to data analysis, drafting or revising the article, have agreed on the journal to which the article will be submitted, gave final approval of the version to be published, and agree to be accountable for all aspects of the work.

Funding

This study was funded by Amgen, Inc.

Disclosure

John H. Page, Ajit A. Londhe, Corinne G. Brooks, Jie Zhang, Carolyn A. Brown, Alvan Cheng, Fang He, Junjie Ma, and Kimberly A. Roehl are employees and stockholders of Amgen, Inc. Katherine B. Carlson, an employee of Moderna, Inc., was formerly an employee of Amgen, Inc. and owns stock in Amgen, Inc. J. Michael Sprafka reports consulting for Amgen, Inc. and owns stock in Amgen, Inc. The authors report no other conflicts of interest in this work.

References

1. Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med*. 2020;382(8):727–733. doi:10.1056/NEJMoa2001017
2. World Health Organization. WHO coronavirus (COVID-19) dashboard. Available from: https://covid19.who.int/. Accessed April 27, 2022.
3. Holshue ML, DeBolt C, Lindquist S, et al. First case of 2019 novel coronavirus in the United States. *N Engl J Med*. 2020;382(10):929–936. doi:10.1056/NEJMoa2001191

4. Jorden MA, Rudman SL, Villarino E, et al.; CDC COVID Response Team. Evidence for limited early spread of COVID-19 within the United States, January-February 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69(22):680–684. doi:10.15585/mmwr.mm6922e1
5. Marik PE, Iglesias J, Varon J, Kory P. A scoping review of the pathophysiology of COVID-19. *Int J Immunopathol Pharmacol*. 2021;35:20587384211048026. doi:10.1177/20587384211048026
6. Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat Med*. 2000;19(9):1141–1164. doi:10.1002/(sici)1097-0258(20000515)19:9<1141::aid-sim479>3.0.co;2-f
7. Centers for Disease Control and Prevention. New ICD-10-CM code for the 2019 novel coronavirus (COVID-19), April 1, 2020; 2020.
8. Kadri SS, Gundrum J, Warner S, et al. Uptake and Accuracy of the diagnosis code for COVID-19 among US hospitalizations. *JAMA*. 2020;324(24):2553–2554. doi:10.1001/jama.2020.20323
9. Bhatt AS, McElrath EE, Claggett BL, et al. Accuracy of ICD-10 diagnostic codes to identify COVID-19 among hospitalized patients. *J Gen Intern Med*. 2021;36(8):2532–2535. doi:10.1007/s11606-021-06936-w
10. Crabb BT, Lyons A, Bale M, et al. Comparison of international classification of diseases and related health problems, tenth revision codes with electronic medical records among patients with symptoms of coronavirus disease 2019. *JAMA Netw Open*. 2020;3(8):e2017703. doi:10.1001/jamanetworkopen.2020.17703
11. Rodgers A, MacMahon S. Systematic underestimation of treatment effects as a result of diagnostic test inaccuracy: implications for the interpretation and design of thromboprophylaxis trials. *Thromb Haemost*. 1995;73(2):167–171. doi:10.1055/s-0038-1653746
12. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.
13. Vinh DB, Zhao X, Kiong KL, et al. Overview of COVID-19 testing and implications for otolaryngologists. *Head Neck*. 2020;42(7):1629–1633. doi:10.1002/hed.26213
14. Zou L, Ruan F, Huang M, et al. SARS-CoV-2 viral load in upper respiratory specimens of infected patients. *N Engl J Med*. 2020;382(12):1177–1179. doi:10.1056/NEJMc2001737
15. Wang W, Xu Y, Gao R, et al. Detection of SARS-CoV-2 in different types of clinical specimens. *JAMA*. 2020;323(18):1843–1844. doi:10.1001/jama.2020.3786
16. Islek A, Balci MK. Analysis of factors causing false-negative real-time polymerase chain reaction results in oropharyngeal and nasopharyngeal swabs of patients with COVID-19. *Ear Nose Throat J*. 2021;145561321996621. doi:10.1177/0145561321996621
17. Barberi C, Castelnuovo E, Dipasquale A, et al. Bronchoalveolar lavage in suspected COVID-19 cases with a negative nasopharyngeal swab: a retrospective cross-sectional study in a high-impact Northern Italy area. *Intern Emerg Med*. 2021;16(7):1857–1864. doi:10.1007/s11739-021-02714-y
18. van Kampen JJA, van de Vijver D, Fraaij PLA, et al. Duration and key determinants of infectious virus shedding in hospitalized patients with coronavirus disease-2019 (COVID-19). *Nat Commun*. 2021;12(1):267. doi:10.1038/s41467-020-20568-4
19. Lee SH. Testing for SARS-CoV-2 in cellular components by routine nested RT-PCR followed by DNA sequencing. *Int J Geriatr Rehabil*. 2020;2:69–96.
20. Roy S. Physicians' dilemma of false-positive RT-PCR for COVID-19: a case report. *SN Compr Clin Med*. 2021;1–4. doi:10.1007/s42399-020-00655-9
21. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford; New York: Oxford University Press; 2003.