Accuracy of Computable Phenotyping Approaches for SARS-CoV-2 Infection and COVID-19 Hospitalizations from the Electronic Health Record

Rohan Khera MD MS[1,2$], Bobak J. Mortazavi PhD[2,3$], Veer Sangha[1], Frederick Warner PhD[2], H. Patrick Young PhD[2], Joseph S. Ross MD MHS[2,4,5], Nilay D. Shah PhD[6], Benjamin D. Pollock PhD[6,7], Karen Wang MD MHS[8,9], Cynthia A. Brandt MD MPH[9,10], Zhenqiu Lin PhD[1,2], Harlan M. Krumholz MD SM[1,2,5], Wade L. Schulz MD PhD[2,9,11*]

[1]Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT

[2]Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, CT

[3]Computer Science & Engineering, Texas A&M University, College Station, TX

[4]Section of General Internal Medicine, Yale School of Medicine, New Haven, CT

[5]Department of Health Policy and Management, Yale School of Public Health, New Haven, CT

[6]Division of Health Care Delivery Research, Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN

[7]Department of Quality, Experience, and Affordability, Mayo Clinic, Rochester, MN

[8]Equity Research and Innovation Center, General Internal Medicine, Yale School of Medicine, New Haven, CT

[9]Center for Medical Informatics, Yale School of Medicine, New Haven, CT

[10]VA Connecticut Healthcare System, West Haven, CT

[11]Department of Laboratory Medicine, Yale School of Medicine, New Haven, CT

[$]Contributed equally as co-first authors

*Corresponding author: Wade L. Schulz, MD, PhD; 55 Church St Suite 804, New Haven CT 06510; Tel: 203-823-8308; E-mail: wade.schulz@yale.edu

ABSTRACT

Objective: Real-world data, including administrative claims and electronic health record (EHR) data, have been critical for rapid-knowledge generation throughout the COVID-19 pandemic. Many studies relied on these data to identify cases and ascertain outcomes., commonly using diagnostic codes. However, to ensure high-quality results are delivered to guide clinical decision making, guide the public health response, and characterize the response to interventions, it is essential to establish the accuracy of these approaches for case identification of infections and hospitalizations.

Methods: Real-world EHR data were obtained from the clinical data warehouse and computational health platform at a large academic health system that includes 5 regional hospitals in Connecticut and Rhode Island and their associated ambulatory practices. Demographic information, diagnosis codes, SARS-CoV-2 nucleic acid and antigen testing results, and visit data including discharge disposition were obtained from our OMOP common data model for all patients with either a positive SARS-CoV-2 test or ICD-10 diagnosis of COVID-19 (U07.1) between April 1, 2020 and March 1, 2021. Various computable phenotype definitions using combinations of test results and diagnostic codes were evaluated for their accuracy to identify SARS-CoV-2 infection and COVID-19 hospitalizations. The association with each phenotype was further compared with case volumes and, for hospitalizations, in-hospital mortality. We conducted a quantitative assessment with a manual chart review for a sample of 40 patients who had discordance between diagnostic code and laboratory result findings.

Results: There were 69,423 individuals with either a diagnosis code or a laboratory diagnosis of a SARS-CoV-2 infection. Of these, 61,023 individuals had a principal or a secondary diagnosis code for COVID-19 and 50,355 had a positive SARS-CoV-2 PCR or antigen test. Among those with a positive PCR, 38,506 (76.5%) also had a principal and 3449 (6.8%) a secondary diagnosis of COVID-19, but 8400 (16.7%) had no COVID-19 diagnosis in the medical record. Moreover, of the 61,023 patients who had a COVID-19 diagnosis, 19,068 (31.2%) did not have

a positive laboratory test for SARS-CoV-2 in the EHR. In a manual chart review of this sample of patients, we found that these many had a COVID-19 diagnosis code added during healthcare encounters related to asymptomatic testing, either as part of a screening program or following exposure, but with negative subsequent test results. The positive predictive value (precision) and sensitivity (recall) of a COVID-19 diagnosis in the medical record for a positive SARS-CoV-2 PCR were 68.8% and 83.3%, respectively. Further, among 5,109 patients who were hospitalized with a principal diagnosis of COVID-19, 4843 (94.8%) had a positive SARS-CoV-2 PCR or antigen test within the 2 weeks preceding hospital admission or during hospitalization. In a random sample of 10 without a positive test during the index hospitalization selected for manual chart review, 7 (70.0%) had been tested at an outside laboratory before admission and the remaining had a strong clinical suspicion for COVID-19. In addition, 789 hospitalizations had a secondary diagnosis of COVID-19, of which 446 (56.5%) had a principal diagnosis that was consistent with severe clinical manifestation of COVID-19 (e.g., sepsis or respiratory failure). Compared with the cohort that had a principal diagnosis of COVID-19, those with a secondary diagnosis more frequently male and White and had more than 2-fold higher in-hospital mortality (13.2% vs 28.0%, P<0.001).

Conclusions: In a large integrated health system, COVID-19 diagnosis codes were not adequate for case identification and epidemiological surveillance of SARS-CoV-2 infection. In contrast, a principal diagnosis of COVID-19 diagnosis consistently identified hospitalized patients with the disease but missed nearly 10% of cases that presented with more severe manifestations of disease and had over 2-fold higher mortality. Data from the EHR can provide additional data elements compared to administrative claims alone, such as laboratory testing results, that can be used to in conjunction with diagnostic codes to create more fine-tuned phenotypes that are designed for specific analytical use cases.

BACKGROUND

The COVID-19 pandemic has led to the rapid adoption of real-world evidence to guide the treatment of and the public health response to a novel pathogen.[1-5] The identification of both SARS-CoV-2 infection and COVID-19 hospitalization is of current clinical and regulatory importance, given the need for case identification for epidemiologic surveillance to track the infections, mortality, and vaccine effectiveness. Similarly, clinical predictive models that rely on appropriate case classification and studies that track the long-term effects of SARS-CoV-2 infection may be biased if case definitions are imprecise or capture only subsets of individuals infected with SARS-CoV-2. Administrative data represent a widely available real-world data source to monitor COVID-19 cases and hospitalizations using diagnosis codes.

Administrative data are generated from billing claims and can be used to assess changes in disease epidemiology, hospitalization rates, and patient outcomes on a large scale.[6-10] However, reliance on claims alone may lead to erroneous inferences, as has been shown for other conditions.[11 12] To ensure that high-quality data guide national policy and biomedical research, there is a need to evaluate the accuracy of the diagnostic code-based approaches used to define cases of SARS-CoV-2 infection and hospitalization.

The adoption of health information technology systems has positioned health systems to allow such validation by incorporating more detailed clinical data from the electronic health record (EHR) with diagnosis codes, which allows for the development of more robust computable phenotypes.[13-18] While there remain concerns about the use of EHR data given potential inaccuracies in data and risk for missing data, these issues are also reflected in data derived from them, such as administrative databases. As such, the EHR represents a potential advance over the use of administrative data alone for both case identification and phenotype validation.

In this study from a large health system with academic and community-based practices, we evaluate the accuracy of various computable phenotypes to identify SARS-CoV-2 infection

and COVID-19 hospitalizations based on diagnostic codes and laboratory testing results extracted from the EHR and assess the impact of cohort selection on the identified mortality rates in those hospitalized for COVID-19.

METHODS

Data Source

We used EHR derived data from Yale New Haven Health System, a single academic health system consisting of 5 distinct hospital delivery networks and associated outpatient clinics located in Connecticut and Rhode Island. Data from our EHR clinical data warehouse were transformed into the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) using our computational health platform.[13 19] We used a versioned extract of the OMOP data from March 3, 2021 and analyzed testing and discharge information from April 1, 2020, when the COVID-19 specific diagnosis was introduced, through March 1, 2021. Admissions were limited to those with a visit start date prior to January 31, 2021 to allow for a majority of those admitted to reach discharge based on our prior data which defined median length of stay for discharge patients.[3]

The study was approved by the Yale University Institutional Review Board (IRB # 2000027747).

Cohort Definitions

SARS-CoV-2 infection: We defined two strategies to identify SARS-CoV-2 infection from the EHR spanning all healthcare settings, the first based on diagnostic codes and the second based on laboratory testing. Our first approach relied on the identification of the specific COVID-19 *International Classification of Diseases-10th Edition-Clinical Modification (ICD-10-CM)* diagnosis code U07.1 within the clinical record. We used the ICD-10-CM code as opposed to the corresponding SNOMED codes in the standard OMOP vocabulary given the wider use of ICD-

10-CM in administrative data. The U07.1 code was used to define SARS-CoV-2 when used either as (1) a principal diagnosis, or a (2) a principal or a secondary diagnosis of COVID-19 during any healthcare encounter. The principal diagnosis was defined based on the standard OMOP condition status concept code, 32902.[20]

The two diagnosis-based phenotyping strategies were compared to the second approach which was based on the presence of a positive SARS-CoV-2 PCR or antigen test to identify individuals who had documented infection, with manual chart abstraction of samples drawn from discordant subsets to assess the reason for differences. We supplemented this assessment to include potentially related but non-specific diagnoses for severe acute respiratory syndrome (SARS) or coronavirus disease (COVID-19-related diagnoses) based on a subset of codes identified within the National COVID Cohort Collaborative (N3C) phenotype (see **eTable 1** in the Online Supplement).[21]

COVID-19 hospitalization: We defined COVID-19 hospitalizations using two strategies. The first identified all hospitalizations with a principal diagnosis of COVID-19 (U07.1). In addition, we defined a second strategy that included individuals with a secondary diagnosis of COVID-19, but with a clinical presentation that was consistent with severe manifestations of COVID-19 defined by a principal diagnosis for acute respiratory failure, pneumonia or sepsis. These approach focuses on hospitalizations that were due to COVID-19 rather than incidentally associated with a positive test for the disease during screening. The principal diagnoses used in this approach are included in **eTable 2**. There was only a single hospitalization with a diagnosis code J12.82 that has been suggested to identify COVID-19 pneumonia,[21] and was not included in the analysis. Similar to the diagnosis-based infection identification, diagnosis codes were compared against positive SARS-CoV-2 testing 2 weeks before hospitalization through any time before hospital discharge.

Study Covariates

We defined key demographic characteristics for individuals, including age, sex, race and ethnicity. Age was defined as completed years on the day of admission, computed from their date of birth. Sex, race and ethnicity were self-reported. To evaluate the effect of coding strategies on case identification among racial and ethnic minorities, we combined racial/ethnic groups into mutually exclusive groups of Hispanic, non-Hispanic White, non-Hispanic Black, and other race/ethnicity groups.[22 23]

Study Outcome

Among patients hospitalized with COVID-19, we evaluated differences in mortality across case identification strategy. Mortality was defined based on the discharge disposition of the first COVID-19 hospitalization. Consistent with other studies,[24-26] we used a composite endpoint of in-hospital mortality, transfer to inpatient hospice or discharge to facility or home-based hospice to define our outcome of early mortality.

Statistical Analyses

We compared difference in demographic characteristics using the chi-square test for categorical variables and t-test for continuous variables. To assess the performance of COVID-19 diagnoses accurately identifying cases of SARS-CoV-2 infection, we assessed 3 key performance measures, precision (positive predictive value), recall (or sensitivity), and area under the precision recall curve (AUPRC).[27 28]

Inpatient hospitalizations were identified the diagnostic strategies outlined above. We explicitly identified the proportion of patients in the two COVID-19 hospitalization groups, those with a principal diagnosis of COVID-19 and those with a secondary diagnosis of COVID-19. We evaluated the implications of case identification strategies on representativeness of the

population through a comparison of demographic features and in-hospital mortality among patients with our two COVID-19 hospitalization phenotypes.

Analyses were conducted using Spark 2.3.2, Python 3.6.9, and R 3.8. All statistical tests were 2-tailed with a level of significance set at 0.05.

Manual Chart Abstraction and Validation

Manual chart abstraction was conducted by 2 clinicians independently (RK and WLS) and focused on a sample of randomly selected charts where the diagnosis codes were discordant from laboratory results. For SARS-CoV-2 infections, 10 patient charts were randomly selected from each of the following categories: (1) principal diagnosis of COVID-19, but negative laboratory diagnosis, (2) secondary diagnosis of COVID-19 but negative laboratory diagnosis, and (3) a positive laboratory diagnosis of SARS-CoV-2 without a corresponding diagnosis code. Furthermore, 10 charts were selected for patients who were hospitalized with a principal diagnosis COVID-19 and negative laboratory results, and the clinical documentation was qualitatively reviewed to evaluate the reason for the discrepancy.

RESULTS

SARS-CoV-2 Testing and Diagnosis Rates

There were 69,423 individuals with either a diagnosis of COVID-19 or a positive PCR for SARS-CoV-2 infection across care settings between April 1, 2020 and March 1, 2021. The mean age of patients was 46.0 (±22.4) years and 45.0% of patients were men. Nearly one fourth of patients were of Hispanic ethnicity (22.8%), 57.6% of patients had a recorded race of White and 15.2% were Black (**Table 1**).

Computable Phenotype Accuracy for SARS-CoV-2 Infection

Of the 69,423 individuals included in our total cohort, 51,540 (74.2%) had a principal diagnosis of COVID-19 in the EHR, 61,023 (87.9%) had a principal or a secondary diagnosis, and 50,355 (72.7%) had a positive SARS-CoV-2 PCR or antigen test. There were consistent differences in number of SARS-CoV-2 infections based on the diagnosis and laboratory-based phenotyping strategies throughout the study period, with diagnostic codes being more common than positive laboratory test findings (**Figure 1**). Similar patterns were observed for non-specific coronavirus diagnoses (**eTable 3** in the Online Supplement).

Bidirectional discordance was noted amongst the phenotyping strategies. Of the 50,355 patients with a positive laboratory test for SARS-CoV-2, only 38,506 (76.5%) had a principal diagnosis and an additional 3449 (6.8%) had a secondary diagnosis of COVID-19. The remaining 8400 (16.7%) had no COVID-19 diagnosis recorded as a principal or secondary diagnostic code within the chart (**Figure 2**). Moreover, there were 19,068 patients (31.2%) who had a principal or a secondary COVID-19 diagnosis without a positive lab test for SARS-CoV-2. The characteristics of patients in these groups are included in Table 1.

In manual chart review of a random sample of 30 cases, all patients with a diagnosis without a positive SARS-CoV-2 identified individuals who had a healthcare visit for SARS-CoV-2 testing but with a subsequent negative laboratory test.

Strategies that identify SARS-CoV-2 infection based on principal diagnosis of COVID-19 had a precision (or positive predictive value) of 74.7% (95% CI, 74.3% to 75.1%) and a recall (or sensitivity) of 76.5% (95% CI, 76.1% to 76.8%), with an AUPRC of 0.17. Inferring infection with any diagnosis code for COVID-19 had worse performance characteristics (precision: 68.8% [68.4% to 69.1%]), recall: 83.3% [83.0% to 83.6%]), and AUPRC 0.12).

There were significant differences in concordance between phenotyping strategies during the study period (**Figure 3**). Among patients with either a diagnosis code for COVID-19 or a laboratory diagnosis of SARS-CoV-2, 51% of patients had both a diagnosis code and a positive laboratory test between April and August 2020, but 65% of patients had both present

between September 2020 and March 2021 (P<0.001). There were significant differences across racial and ethnic groups (**Figure 4A**). Among Hispanics and non-Hispanic Black patients, 69.5% and 68.7% patients, respectively had both a diagnosis of COVID-19 and a positive laboratory test, compared with 54.5% of non-Hispanic White patients (P<0.001). There was also a significant difference by sex, with more women compared with men had a concomitant diagnosis code and positive laboratory test (61.4% vs 59.1%, P<0.001) (**Figure 4B**).

Computable Phenotype Accuracy for COVID-19 Hospitalization

Based on visit start date, there were a total of 5,555 discharges based on our overall phenotyping strategy between April 1, 2020 and January 31, 2021. Of these, 5,109 (92.0%) discharges had a principal diagnosis of COVID-19 and the remaining 446 had a primary diagnosis that could be associated with a severe presentation with a secondary diagnosis of COVID-19 on the same visit. Finally, there were 343 individuals who had a secondary, but not primary, diagnosis of COVID-19 which were excluded from analysis as these may indicate either incidental findings or hospital-acquired infections.

A vast majority of patients had a positive SARS-CoV-2 PCR or antigen test during the hospitalization, across both patients hospitalized with a COVID-19 primary diagnosis (94.8%, n=4843) or a secondary diagnosis (91.9%, n=410). A manually abstracted sample of 10 charts of hospitalized individuals without a positive laboratory test but a principal diagnosis of COVID during the test identified that 7 of these patients had a positive COVID-19 test at another healthcare facility prior to presentation and 3 had a strong clinical suspicion for COVID-19 but a negative PCR test.

Relationship between COVID-19 Hospitalization Phenotype Definition and In-Hospital Mortality Rate

The in-hospital mortality rate for those hospitalized with a principal diagnosis of COVID-19 was 13.2% (675 of 5,109), which was significantly lower than those with a secondary diagnosis of COVID-19 and related primary diagnosis code for sepsis or respiratory failure, who had an in-hospital mortality rate that was nearly double (28.0%, 125 of 446, P<0.001) (**Table 2**). Those with a COVID-19 primary diagnosis were less frequently male (50.9% vs 59.6%, P <0.001) and were more frequently Black (20.9% vs 17.3%, P = 0.02).

DISCUSSION

In a large integrated health system with 5 hospital care delivery networks and their outpatient clinics, COVID-19 diagnosis codes were not adequate for case identification and epidemiological surveillance of SARS-CoV-2 infection. In contrast, a principal diagnosis of COVID-19 diagnosis consistently identified patients with the disease, though would miss nearly 10% of cases who presented with a clinical disease consistent with COVID-19, but only with a secondary diagnosis of COVID-19. A focus on primary diagnosis of COVID-19 to identify hospitalizations is likely to miss this high-risk group, who had an over 2-fold higher mortality than patients with principal diagnosis of COVID-19. In an assessment spanning outpatient and inpatient settings, there was substantial discordance in a diagnosis code and a laboratory diagnosis of SARS-CoV-2 infection with 1 in 3 patients with a COVID-19 diagnosis without a SARS-CoV-2 infection based on laboratory testing, and 1 in 5 with a positive SARS-Cov-2 PCR or antigen test without any corresponding diagnosis codes.

To our knowledge, in the largest study leveraging the EHR, rather than administrative claims, to evaluate COVID-19 phenotypes across outpatient and inpatient healthcare settings across an entire year. Moreover, in addition to identifying the accuracy of phenotypic definitions, we evaluate the association of phenotype definitions on inferred short-term outcomes. Previous investigations highlighted that the COVID-19 diagnosis code, U07.1 was rapidly incorporated into the workflow of US hospitals in early 2020,[29] and among hospitalized patients, with a high

sensitivity and specificity of the code for laboratory confirmed disease. However, the study was limited in using data only through May 2020, with only 4965 SARS-CoV-2 positive laboratory tests, with the accuracy measures drive by the 89.6% of the cohort that did not have either a positive test or a diagnosis code for COVID-19.[29] We confirm that an inpatient diagnosis of COVID-19 has retained a large positive predictive value for clinical COVID-19. However, we found significant heterogeneity in outcomes based on whether COVID-19 was included as a principal or a secondary diagnosis. Moreover, the study evaluating COVID-19 diagnosis codes focused exclusively among hospitalized patients and with did not evaluate the role of diagnosis codes in case surveillance.

A notable strength of our approach based on the EHR is the ability to gain additional insights into the mechanism behind discordance between diagnosis codes and laboratory results through manual chart review. This allowed us to carefully define a cohort of patients of COVID-19 consistent with its clinical definition. Moreover, it also allowed us to exclude patients who do not have COVID-19 despite a secondary diagnosis as well as a positive SARS-CoV-2 test, as during the pandemic many clinical representations (such as presentations for trauma), did in fact have an incidental positive SARS-CoV-2 screen without clinical disease.[3] Moreover, our study population includes a large racially and ethnically diverse group, allowing our study to uncover patterns of differential performance of diagnosis codes for case surveillance in racial/ethnic minority groups, in whom the presence of diagnosis codes more accurately aligned with laboratory confirmed disease.

There are many possible reasons for the incorrect classification of infections by diagnosis codes. Many studies have shown the apparent inaccuracy of various EHR data elements, such as the clinical history and problem list.[30 31] Clinical uncertainty related to a diagnosis, potential stigma associated with the addition of a diagnosis to the medical record, clinical workflows that do not promote the capture of structured data elements, and miscoded diagnoses can all impact the ability to define a digital phenotype that accurately identifies

patients.[30 32 33] Moreover, diagnosis codes are often included when evaluating a suspected condition and may be misconstrued as a proof of the diagnosis, particularly in data captured in real-time. Our study highlights the value of health information systems in disease monitoring through logical cohort definitions, which can be explicitly confirmed across different data streams. Moreover, our work complements the continuous monitoring and validation of claim-based diagnoses that is essential for valid inference from a large number of investigations that rely on these ubiquitous data. A unique aspect of our approach is the availability of all EHR data in the OMOP CDM that is updated daily, allowing rapid and serially updated assessment of such cohort definitions and disease surveillance over the course of the pandemic, allowing the most up-to-date assessment of specific patient populations in a rapidly evolving condition and their outcomes.

Our study has certain limitations. First, while we focus on a broad interconnected health system and affiliated laboratories and receive testing information from laboratories that are connected to the Epic EHR, it is possible that some of the non-participating laboratory data are not available from testing in the outpatient setting. However, in manual chart review of a sample of patients with a diagnosis of COVID-19 without a reported positive PCR or antigen test by 2 clinicians, all such records were for patients undergoing SARS-CoV-2 testing with the diagnosis assigned for the clinical or laboratory encounter to obtain the test. Second, we cannot account for differences in coding practices at other institutions. However, our study that includes a large integrated multi-hospital health system, which increases the generalizability of our observations. Moreover, a larger variation in coding of diagnoses for SARS-CoV-2 infection surveillance would further highlight the lack of reliability of the measures.

CONCLUSIONS

Among hospitalized patients with COVID-19 rate of mortality to be significantly higher in hospitalized patients who had a principal diagnosis code consistent with severe manifestations

of COVID-19 compared to those who had a specific principal diagnosis of COVID-19. Moreover, the use of COVID-19 diagnosis codes alone frequently misclassifies individuals with a SARS-CoV-2 infection and may not be an adequate strategy for case surveillance. Data from the EHR can provide additional data elements such as laboratory testing results that can be used to in conjunction with diagnostic codes to create more fine-tuned phenotypes that are designed for specific analytical use cases.

Funding: None

Disclosures: H.M.K. works under contract with the Centers for Medicare & Medicaid Services to support quality measurement programs, was a recipient of a research grant from Johnson & Johnson, through Yale University, to support clinical trial data sharing; was a recipient of a research agreement, through Yale University, from the Shenzhen Center for Health Information for work to advance intelligent disease prevention and health promotion; collaborates with the National Center for Cardiovascular Diseases in Beijing; receives payment from the Arnold & Porter Law Firm for work related to the Sanofi clopidogrel litigation, from the Martin Baughman Law Firm for work related to the Cook Celect IVC filter litigation, and from the Siegfried and Jensen Law Firm for work related to Vioxx litigation; chairs a Cardiac Scientific Advisory Board for UnitedHealth; was a member of the IBM Watson Health Life Sciences Board; is a member of the Advisory Board for Element Science, the Advisory Board for Facebook, and the Physician Advisory Board for Aetna; and is the co-founder of HugoHealth, a personal health information platform, and co-founder of Refactor Health, a healthcare AI-augmented data management company. W.L.S. was an investigator for a research agreement, through Yale University, from the Shenzhen Center for Health Information for work to advance intelligent disease prevention and health promotion; collaborates with the National Center for Cardiovascular Diseases in Beijing; is a technical consultant to HugoHealth, a personal health information platform, and co-founder of Refactor Health, an AI-augmented data management platform for healthcare; is a

consultant for Interpace Diagnostics Group, a molecular diagnostics company. J.S.R. currently receives research support through Yale University from Johnson and Johnson to develop methods of clinical trial data sharing, from the Medical Device Innovation Consortium as part of the National Evaluation System for Health Technology (NEST), from the Food and Drug Administration for the Yale-Mayo Clinic Center for Excellence in Regulatory Science and Innovation (CERSI) program (U01FD005938); from the Agency for Healthcare Research and Quality (R01HS022882), from the National Heart, Lung and Blood Institute of the National Institutes of Health (NIH) (R01HS025164, R01HL144644), and from the Laura and John Arnold Foundation to establish the Good Pharma Scorecard at Bioethics International. In the past 36 months, NDS has received research support through Mayo Clinic from the Food and Drug Administration to establish Yale-Mayo Clinic Center for Excellence in Regulatory Science and Innovation (CERSI) program (U01FD005938); the Centers of Medicare and Medicaid Innovation under the Transforming Clinical Practice Initiative (TCPI); the Agency for Healthcare Research and Quality (R01HS025164; R01HS025402; R03HS025517; K12HS026379); the National Heart, Lung and Blood Institute of the National Institutes of Health (NIH) (R56HL130496; R01HL131535; R01HL151662); the National Science Foundation; from the Medical Device Innovation Consortium as part of the National Evaluation System for Health Technology (NEST)and the Patient Centered Outcomes Research Institute (PCORI) to develop a Clinical Data Research Network (LHSNet). The other authors do not report any relevant disclosures.

References

1. Haimovich AD, Ravindra NG, Stoytchev S, et al. Development and Validation of the Quick COVID-19 Severity Index: A Prognostic Tool for Early Clinical Decompensation. Ann Emerg Med 2020;**76**(4):442-53 doi: 10.1016/j.annemergmed.2020.07.022.
2. Burn E, You SC, Sena AG, et al. Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study. Nat Commun 2020;**11**(1):5009 doi: 10.1038/s41467-020-18849-z.
3. McPadden J, Warner F, Young HP, et al. Clinical Characteristics and Outcomes for 7,995 Patients with SARS-CoV-2 Infection. medRxiv 2020 doi: 10.1101/2020.07.19.20157305.
4. Hendren NS, de Lemos JA, Ayers C, et al. Association of Body Mass Index and Age With Morbidity and Mortality in Patients Hospitalized With COVID-19: Results From the American Heart Association COVID-19 Cardiovascular Disease Registry. Circulation 2021;**143**(2):135-44 doi: 10.1161/CIRCULATIONAHA.120.051936.
5. Khera R, Clark C, Lu Y, et al. Association of Angiotensin-Converting Enzyme Inhibitors and Angiotensin Receptor Blockers with the Risk of Hospitalization and Death in Hypertensive Patients with Coronavirus Disease-19. J Am Heart Assoc 2021:e018086 doi: 10.1161/JAHA.120.018086.
6. Khera R, Dharmarajan K, Wang Y, et al. Association of the Hospital Readmissions Reduction Program With Mortality During and After Hospitalization for Acute Myocardial Infarction, Heart Failure, and Pneumonia. JAMA Netw Open 2018;**1**(5):e182777 doi: 10.1001/jamanetworkopen.2018.2777.
7. Khera R, Pandey A, Ayers Colby R, et al. Contemporary Epidemiology of Heart Failure in Fee-For-Service Medicare Beneficiaries Across Healthcare Settings. Circulation: Heart Failure 2017;**10**(11):e004402 doi: 10.1161/CIRCHEARTFAILURE.117.004402.
8. Khera R, Wang Y, Nasir K, Lin Z, Krumholz HM. Evaluation of 30-Day Hospital Readmission and Mortality Rates Using Regression-Discontinuity Framework. Journal of the American College of Cardiology 2019;**74**(2):219-34 doi: 10.1016/j.jacc.2019.04.060.
9. Khera R, Wang Y, Bernheim SM, Lin Z, Krumholz HM. Post-discharge acute care and outcomes following readmission reduction initiatives: national retrospective cohort study of Medicare beneficiaries in the United States. BMJ (Clinical research ed.) 2020;**368**:l6831 doi: 10.1136/bmj.l6831.
10. Khera R, Kondamudi N, Zhong L, et al. Temporal Trends in Heart Failure Incidence Among Medicare Beneficiaries Across Risk Factor Strata, 2011 to 2016. JAMA Netw Open 2020;**3**(10):e2022190 doi: 10.1001/jamanetworkopen.2020.22190.
11. Khera R, Spertus JA, Starks MA, et al. Administrative Codes for Capturing In-Hospital Cardiac Arrest. JAMA Cardiol 2017;**2**(11):1275-77 doi: 10.1001/jamacardio.2017.2904.
12. Muscatello DJ, Amin J, MacIntyre CR, et al. Inaccurate ascertainment of morbidity and mortality due to influenza in administrative databases: a population-based record linkage study. PLoS One 2014;**9**(5):e98446 doi: 10.1371/journal.pone.0098446.
13. Schulz WL, Durant TJS, Torre CJ, Jr., Hsiao AL, Krumholz HM. Agile Health Care Analytics: Enabling Real-Time Disease Surveillance With a Computational Health Platform. J Med Internet Res 2020;**22**(5):e18707 doi: 10.2196/18707.

14. Aronsky D, Haug PJ, Lagor C, Dean NC. Accuracy of administrative data for identifying patients with pneumonia. Am J Med Qual 2005;**20**(6):319-28 doi: 10.1177/1062860605280358.
15. Bourgon Labelle J, Farand P, Vincelette C, Dumont M, Le Blanc M, Rochefort CM. Validation of an algorithm based on administrative data to detect new onset of atrial fibrillation after cardiac surgery. BMC Med Res Methodol 2020;**20**(1):75 doi: 10.1186/s12874-020-00953-9.
16. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013;**20**(1):117-21 doi: 10.1136/amiajnl-2012-001145.
17. Peabody JW, Luck J, Jain S, Bertenthal D, Glassman P. Assessing the accuracy of administrative data in health information systems. Medical care 2004;**42**(11):1066-72 doi: 10.1097/00005650-200411000-00005.
18. van Mourik MS, van Duijn PJ, Moons KG, Bonten MJ, Lee GM. Accuracy of administrative data for surveillance of healthcare-associated infections: a systematic review. BMJ Open 2015;**5**(8):e008424 doi: 10.1136/bmjopen-2015-008424.
19. McPadden J, Durant TJ, Bunch DR, et al. Health Care and Precision Medicine Research: Analysis of a Scalable Data Science Platform. J Med Internet Res 2019;**21**(4):e13043 doi: 10.2196/13043.
20. Observational Health Data Sciences and Informatics. ATHENA. https://athena.ohdsi.org/. Accessed March 9, 2021.
21. National COVID Cohort Collaborative. Phenotype Data Acquisition. Available at: https://github.com/National-COVID-Cohort-Collaborative/Phenotype_Data_Acquisition/wiki/Latest-Phenotype. Accessed March 3, 2021.
22. Flanagin A, Executive Managing Editor JatJN, Frey T, et al. The Reporting of Race and Ethnicity in Medical and Science Journals: Comments Invited. JAMA 2021;**325**(11):1049-52 doi: 10.1001/jama.2021.2104.
23. Boyd RW, Lindo EG, Weeks LD, McLemore MR. On Racism: A New Standard For Publishing On Racial Health Inequities | Health Affairs Blog. 2021 doi: 10.1377/hblog20200630.939347.
24. Razavian N, Major VJ, Sudarshan M, et al. A validated, real-time prediction model for favorable outcomes in hospitalized COVID-19 patients. NPJ Digit Med 2020;**3**:130 doi: 10.1038/s41746-020-00343-x.
25. Ogedegbe G, Department of Population Health NYUGSoM, New York, Department of Medicine NYUGSoM, New York, et al. Assessment of Racial/Ethnic Disparities in Hospitalization and Mortality in Patients With COVID-19 in New York City. JAMA Network Open 2021;**3**(12) doi: 10.1001/jamanetworkopen.2020.26881.
26. Petrilli CM, Jones SA, Yang J, et al. Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: prospective cohort study. BMJ 2020;**369**:m1966 doi: 10.1136/bmj.m1966.
27. Sofaer HR, Hoeting JA, Jarnevich CS, McPherson J. The area under the precision-recall curve as a performance metric for rare binary events. Methods in Ecology and Evolution 2019;**10**(4):565-77 doi: 10.1111/2041-210x.13140.

28. Cook J, Ramadas V. When to consult precision-recall curves. Stata J 2020;**20**(1):131-48 doi: 10.1177/1536867x20909693.
29. 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):2553-54 doi: 10.1001/jama.2020.20323.
30. Schulz WL, Young HP, Coppi A, et al. Temporal relationship of computed and structured diagnoses in electronic health record data. BMC Med Inform Decis Mak 2021;**21**(1):61 doi: 10.1186/s12911-021-01416-x.
31. Hartzema AG, Racoosin JA, MaCurdy TE, Gibbs JM, Kelman JA. Utilizing Medicare claims data for real-time drug safety evaluations:is it feasible? Pharmacoepidemiol Drug Saf 2011;**20**(7):684-8 doi: 10.1002/pds.2143.
32. Holmes C, Brown M, Hilaire DS, Wright A. Healthcare provider attitudes towards the problem list in an electronic health record: a mixed-methods qualitative study. BMC Med Inform Decis Mak 2012;**12**(1):127 doi: 10.1186/1472-6947-12-127.
33. Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB. Discordance of databases designed for claims payment versus clinical information systems. Implications for outcomes research. Ann Intern Med 1993;**119**(8):844-50 doi: 10.7326/0003-4819-119-8-199310150-00011.

FIGURES

Figure 1: Cumulative SARS-CoV-2 infection volumes by adjudication strategy across the study period.

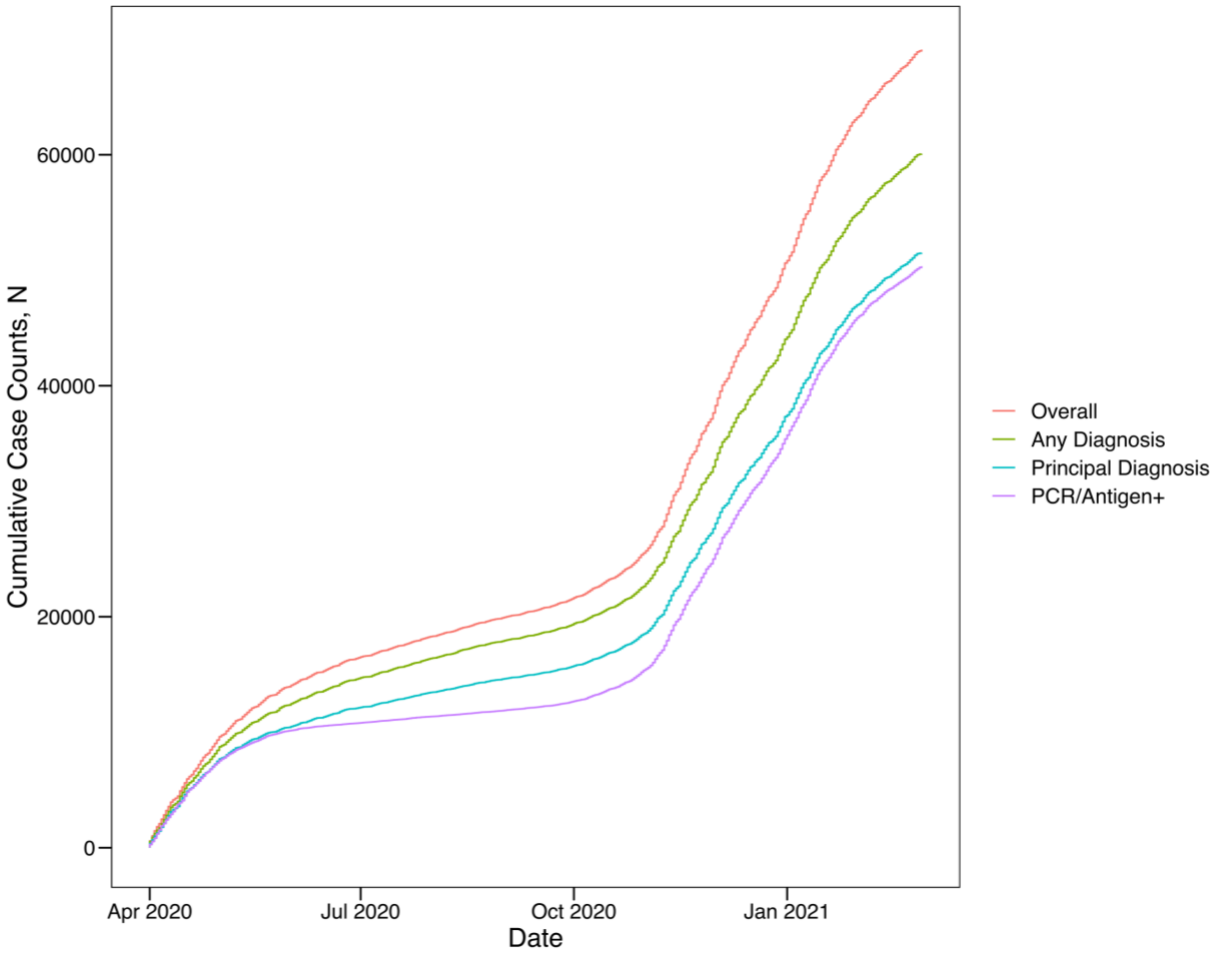


Figure 2: Diagnostic groups for SARS-CoV2 infection across the study period

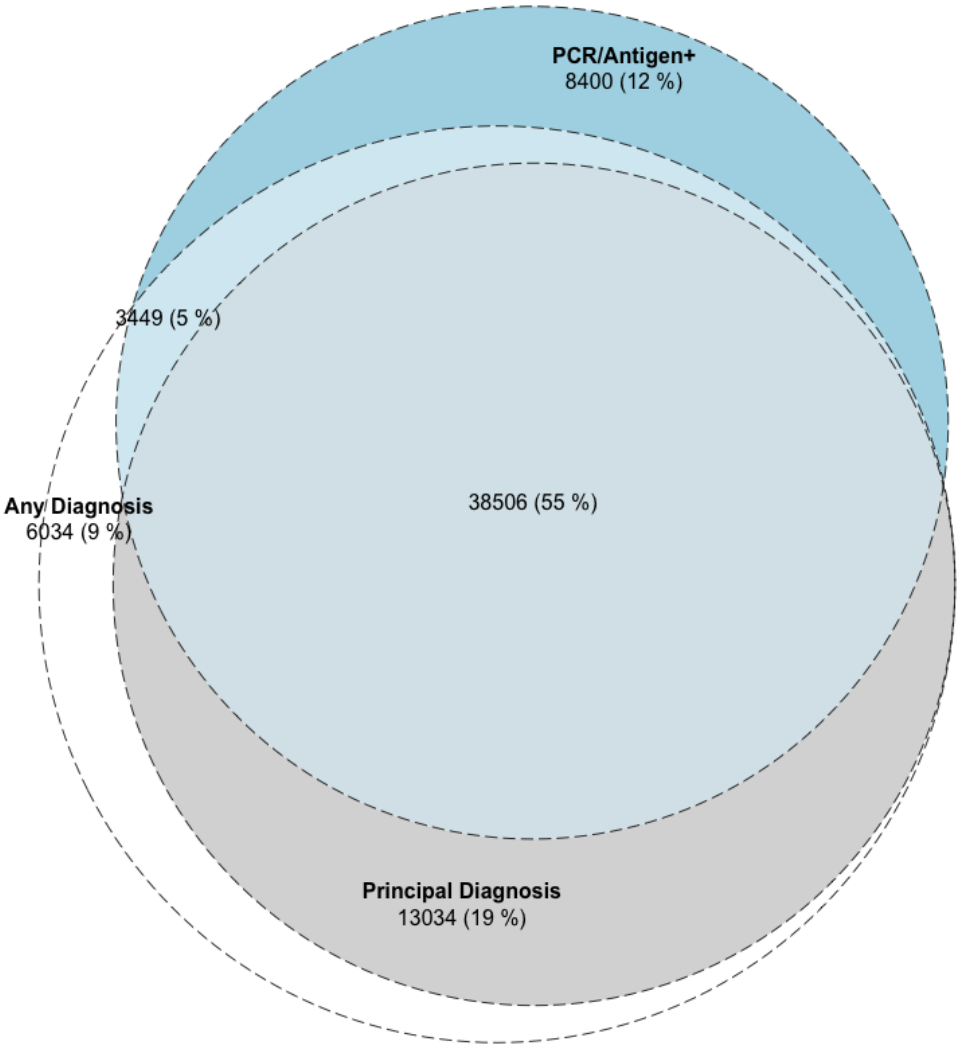


Figure 3: Diagnostic groups for SARS-CoV2 infection across the study period, (A) April 2020 to August 2020, (B) September 2020 to March 2021

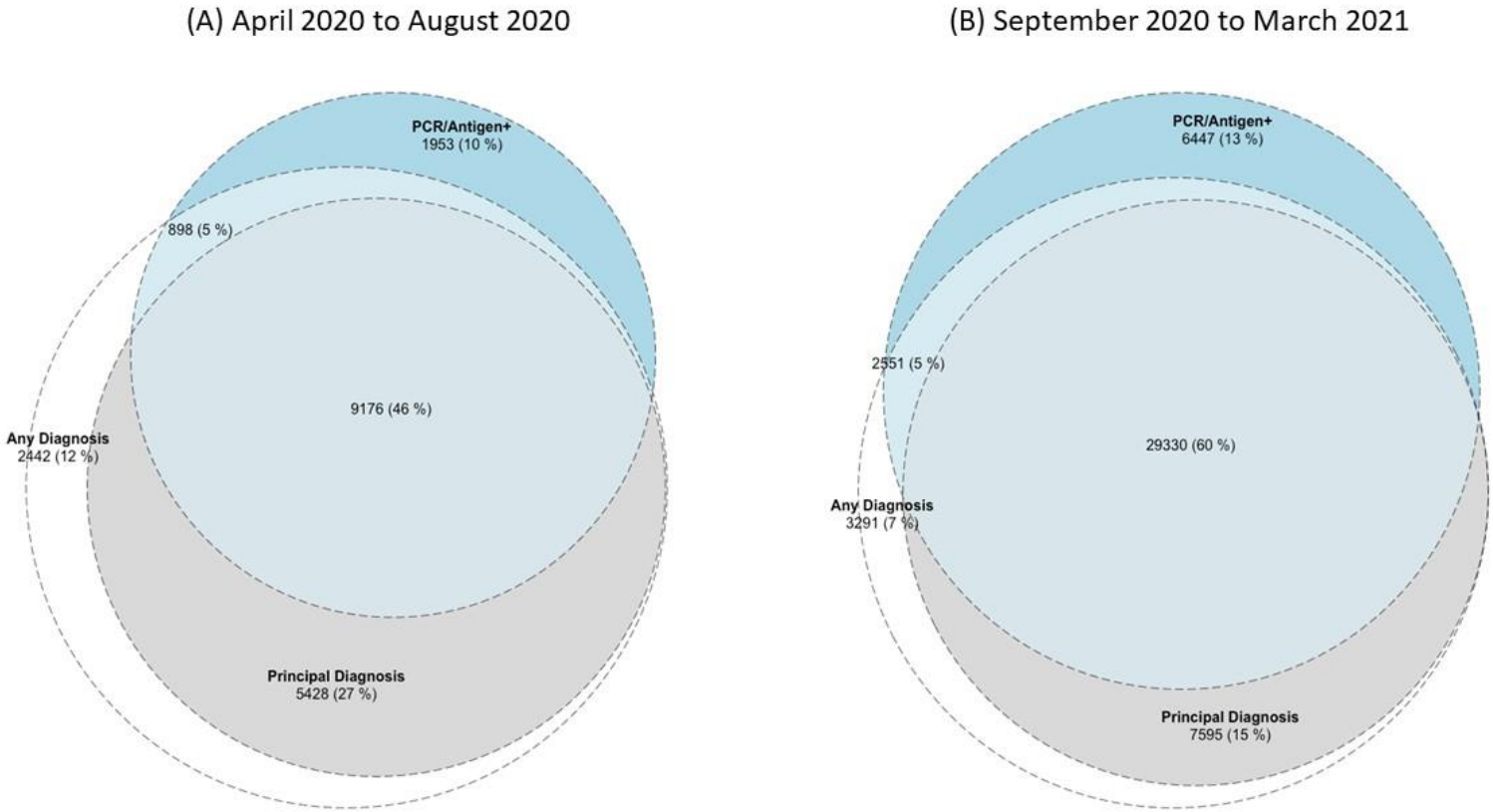
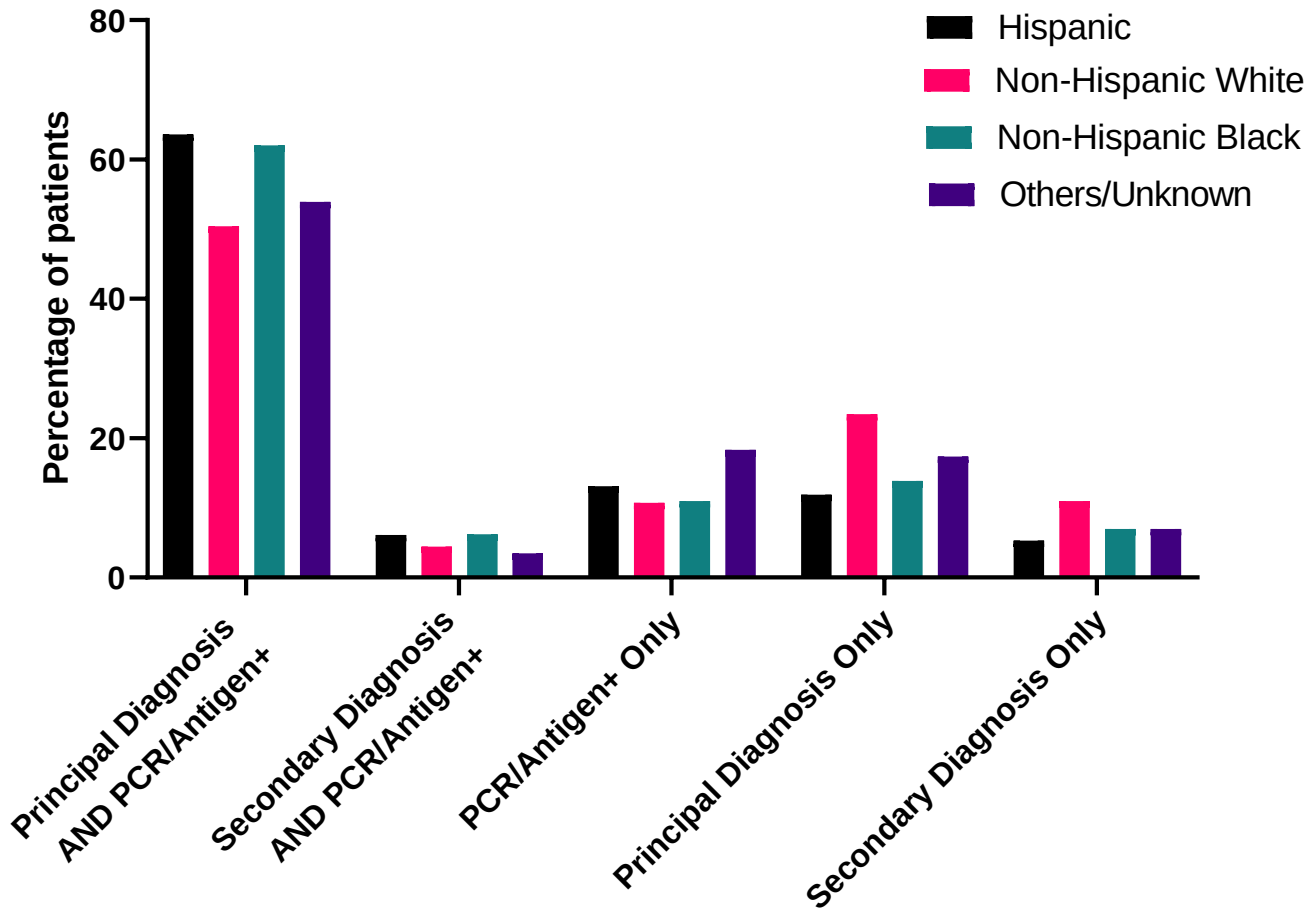


Figure 4: Diagnostic groups for SARS-CoV2 infection by (A) Race/Ethnicity, and (B) Sex.

4(A)



4 (B)

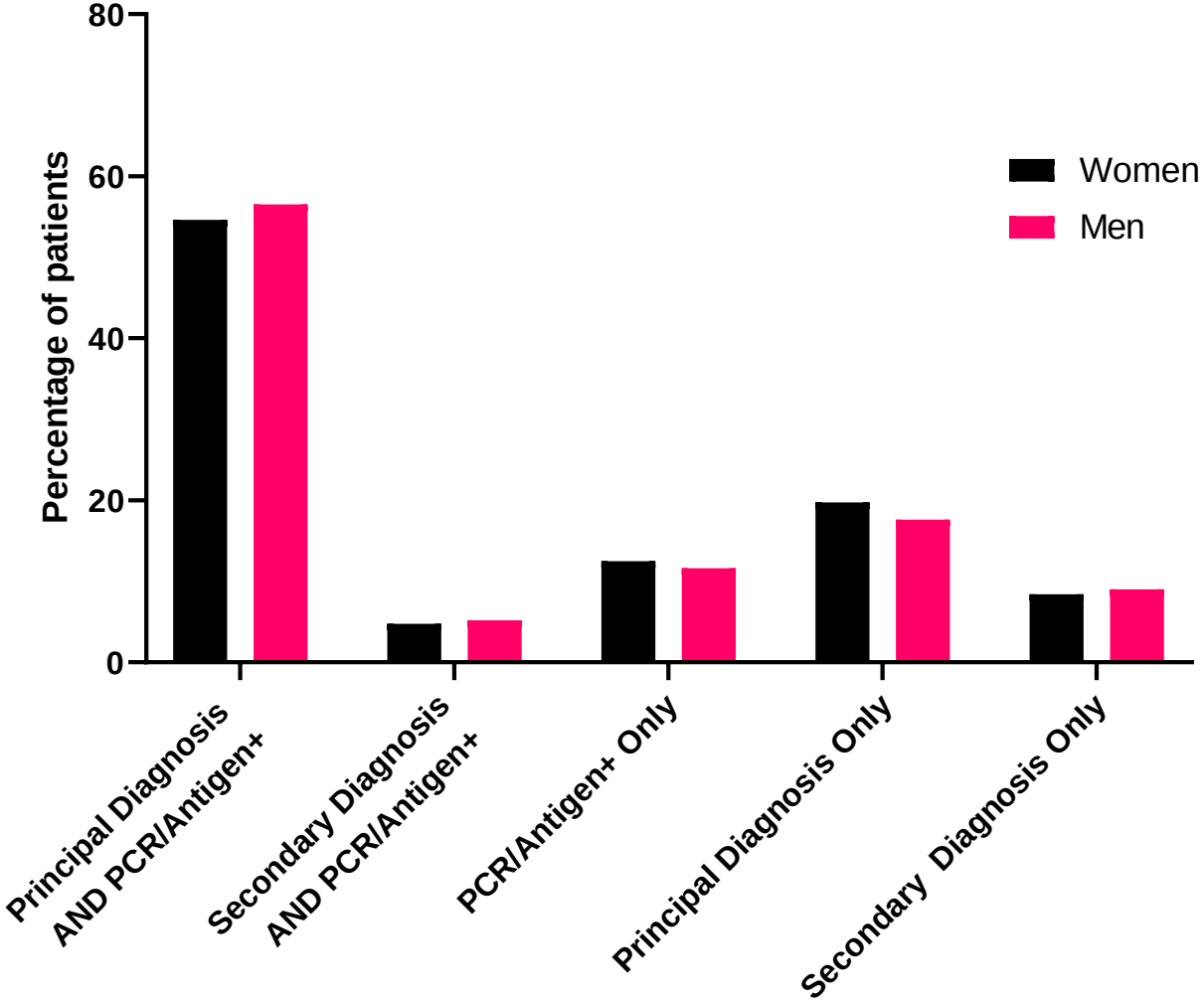

80
60
40
20
0
Percentage of patients
Women
Men
Principal Diagnosis AND PCR/Antigen+
Secondary Diagnosis AND PCR/Antigen+
PCR/Antigen+ Only
Principal Diagnosis Only
Secondary Diagnosis Only

Table 1: Characteristics of patients across mutually exclusive diagnostic groups.

Characteristics	Overall	Principal diagnosis PLUS PCR/Antigen+	Secondary PLUS PCR/Antigen+	Principal diagnosis only	Secondary diagnosis only	PCR/Antigen+ only
Number of Patients	69423	38506	3449	13034	6034	8400
Age (mean (SD))	46.00 (22.44)	44.07 (22.29)	51.15 (23.84)	49.56 (21.25)	52.41 (24.56)	42.59 (20.70)
Men, n (%)						
Race, n (%)	31271 (45.0)	17671 (45.9)	1629 (47.2)	5512 (42.3)	2823 (46.8)	3636 (43.3)
Black						
White	10582 (15.2)	6569 (17.1)	650 (18.8)	1460 (11.2)	732 (12.1)	1171 (13.9)
Asian	39976 (57.6)	20665 (53.7)	1797 (52.1)	8973 (68.8)	4221 (70.0)	4320 (51.4)
Native Hawaiian/Other Pacific Islander	1248 (1.8)	679 (1.8)	53 (1.5)	285 (2.2)	87 (1.4)	144 (1.7)
American Indian or Alaska Native	242 (0.3)	138 (0.4)	13 (0.4)	43 (0.3)	11 (0.2)	37 (0.4)
Other Race	144 (0.2)	73 (0.2)	10 (0.3)	31 (0.2)	10 (0.2)	20 (0.2)
Unknown	11833 (17.0)	7418 (19.3)	789 (22.9)	1421 (10.9)	619 (10.3)	1586 (18.9)
Ethnicity (%)	5398 (7.8)	2964 (7.7)	137 (4.0)	821 (6.3)	354 (5.9)	1122 (13.4)
Hispanic or Latino						
Not Hispanic or Latino	15829 (22.8)	10071 (26.2)	966 (28.0)	1882 (14.4)	838 (13.9)	2072 (24.7)
Unknown	47721 (68.7)	25268 (65.6)	2352 (68.2)	10185 (78.1)	4741 (78.6)	5175 (61.6)

Table 2: Characteristics of hospitalized COVID-19 patients with a principal or secondary diagnosis of COVID-19 (U07.1).

Characteristics	Overall	Principal diagnosis of COVID-19	Secondary diagnosis of COVID-19*
Number of Patients	5555	5109	446
Age (mean (SD))	66.37 (17.59)	66.17 (17.68)	68.63 (16.44)
Men, n (%)	2867 (51.6)	2601 (50.9)	266 (59.6)
Race, n (%)			
Black	1145 (20.6)	1068 (20.9)	77 (17.3)
White	3156 (56.8)	2880 (56.4)	276 (61.9)
Asian	103 (1.9)	96 (1.9)	7 (1.6)
Native Hawaiian/Other Pacific Islander	19 (0.3)	19 (0.4)	0 (0.0)
American Indian or Alaska Native	12 (0.2)	11 (0.2)	1 (0.2)
Other Race	1043 (18.8)	960 (18.8)	83 (18.6)
Unknown	77 (1.4)	75 (1.5)	2 (0.4)
Ethnicity (%)			
Hispanic or Latino	1243 (22.4)	1152 (22.5)	91 (20.4)
Not Hispanic or Latino	4259 (76.7)	3908 (76.5)	351 (78.7)
Unknown	53 (1.0)	49 (1.0)	4 (0.9)
In-hospital mortality/Discharge to Hospice, n (%)	800 (14.4)	675 (13.2)	125 (28.0)

*with a principal diagnosis for respiratory failure, sepsis or pneumonia