cmajOPEN

Research

Validity of hospital diagnostic codes to identify SARS-CoV-2 infections in reference to polymerase chain reaction results: a descriptive study

Cristiano S. Moura PhD, Autumn Neville BA, Fangming Liao MSc, Bijun Wen PhD, Fahad Razak MD MSc, Surain Roberts PhD, Amol A. Verma MD MPhil, Sasha Bernatsky MD PhD

Abstract

Background: In 2020, *International Statistical Classification of Diseases and Related Health Problems, 10th Revision* (ICD-10) codes were created for laboratory-confirmed SARS-CoV-2 infections. We assessed the operating characteristics of ICD-10 discharge diagnostic code U07.1 within the General Medicine Inpatient Initiative (GEMINI).

Methods: GEMINI assembles hospitalization data (including administrative ICD-10 discharge diagnostic codes, laboratory results and demographic data) from hospitals in Ontario, Canada. We studied adults (age ≥ 18 yr) admitted during 2020 and tested at least once for SARS-CoV-2 via polymerase chain reaction (PCR) during (or within 48 h before) hospitalization. With PCR results as the reference standard, we calculated sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) for ICD-10 code U07.1 hospital discharge diagnostic codes. Analyses were stratified by demographic data, calendar period and timing of the first test (within or after 48 h of hospital admission).

Results: In 11 852 hospitalizations with at least 1 SARS-CoV-2 PCR test, 444 (3.7%) were positive. The sensitivity of code U07.1 to identify SARS-CoV-2 infection was 97.8%, specificity was 99.5%, PPV was 88.2% and NPV was 99.9%. Operating characteristics were similar in most stratified analyses, but the specificity and PPV were lower if the first SARS-CoV-2 test was done more than 48 hours after admission.

Interpretation: The sensitivity, specificity, PPV and NPV of code U07.1 were high. This supports using code U07.1 to identify SARS-CoV-2 infection in hospitalization data.

Since the emergence of COVID-19, the global health care community faced unprecedented challenges in identifying and tracking patients infected with the novel coronavirus, SARS-CoV-2. The urgency to contain the spread of the virus prompted the development of efficient and accurate methods for patient identification. During its initial phase, various approaches were employed to identify individuals with suspected or confirmed cases of COVID-19, ranging from clinical diagnosis to laboratory testing. However, the lack of standardized case definitions hindered the comparability of data across different health care systems and regions. In early 2020, the World Health Organization (WHO) released new *International Statistical Classification of Diseases and Related Health Problems, 10th Revision* (ICD-10) codes for the identification of confirmed (U07.1, virus identified) and suspected cases (U07.2, virus not identified) of SARS-CoV-2 infection. These codes were implemented in Canada in April 2020[1] (further additions were introduced early in 2021 to enable more comprehensive data capture, including U07.4 [post-COVID-19 condition] and U07.5 [personal history of COVID-19]).[2]

Reliable, standardized case definitions are needed to make optimal use of routinely collected health data, particularly administrative health care records. A handful of studies have assessed ICD-10 code U07.1 in North America and elsewhere.[3–5] If this new code can reliably identify SARS-CoV-2 infection within health data, it could expedite important research and surveillance activities. Given the potential for variability across different jurisdictions (coding practices, patient populations and other factors), we aimed to confirm this code's validity within a broad patient population with publicly funded health care.

Competing interests: None declared.

This article has been peer reviewed.

Correspondence to: Sasha Bernatsky, sasha.bernatsky@mcgill.ca

***CMAJ Open* 2023 October 24. DOI:10.9778/cmajo.20230033**

Methods

The General Medicine Inpatient Initiative (GEMINI) is a hospital research collaborative collecting clinical and administrative data from hospitals in the province of Ontario, Canada. GEMINI receives discharge diagnosis codes as reported by hospitals. Data include inpatient and emergency department care, including demographic data, administrative discharge diagnostic codes, vital signs, and laboratory test results and imaging.[6,7] In Canada, upon discharge, trained medical clerks at each hospital assign administrative discharge diagnostic codes (1 most responsible diagnosis and up to 25 additional codes).[8] We studied all adults (age ≥ 18 yr) admitted to 1 of 7 GEMINI participating hospitals between January and December 2020 with at least 1 SARS-CoV-2 polymerase chain reaction (PCR) test at admission (or the 48 h preceding) or during hospitalization. This was the period when complete data were available in all the GEMINI hospitals included in the analysis. It also aligned with our intention to examine the operating characteristics of the new ICD-10 code in the context of its early implementation. During this period in Ontario, rapid antigen testing was generally unavailable for the general population, and confirmation of suspected SARS-CoV-2 infection by PCR test was mandated by public health authorities. The hospitals were located in the Greater Toronto Area (Mount Sinai, Sunnybrook Hospital, St. Michael's Hospital, Toronto Western Hospital, Toronto General Hospital) and Mississauga (Trillium Health Partners Credit Valley and Mississauga Hospitals).

Design, index and reference standard tests, and clinical variables

We characterized demographic data (sex, age and urban v. rural residence as per Statistics Canada[9]).

Methods for estimating or comparing measures of diagnostic accuracy

To evaluate the performance of the U07.1 code using PCR test results as the reference standard, we identified all SARS-CoV-2 PCR tests performed 48 hours before admission and all tests performed during hospitalization. Readmissions were treated as independent observations, meaning that any tests performed in previous encounters were generally not accounted for in the current admission. We assumed cases as laboratory-confirmed (SARS-CoV-2 identified) if at least 1 test was positive in the observation period, and we assumed a confirmed noncase (no SARS-CoV-2 identified) if there were no positive PCR tests (and at least 1 negative test result). We then defined a true positive as laboratory-confirmed SARS-CoV-2, with discharge code U07.1; a false negative as laboratory-confirmed SARS-CoV-2, without discharge code U07.1; a false positive as no SARS-CoV-2 identified, with discharge code U07.1; and a true negative as no SARS-CoV-2 identified, without discharge code U07.1. We then calculated sensitivity as the number of true positives divided by the sum of true positives and false negatives, specificity as the number of true negatives divided by the sum of true negatives and false positives, positive predictive value (PPV) as the number of true positives divided by the sum of true positives and false positives, and negative predictive value (NPV) as the number of true negatives divided by the sum of true negatives and false negatives. These estimates and their corresponding 95% CI were computed using the epiR R package.[10]

Data analysis

Stratified analyses were carried out for sex and age (< 50 yr, 50–75 yr and > 75 yr), urban versus rural residence, calendar period of admission (January–April, May–August and September–December) and timing of first PCR test (before admission, within 24 h of the admission, between 24 and 48 h of the admission, or after 48 h of admission).

All analyses in this paper were performed using R version 4.1.2.[10]

Ethics approval

The McGill University Research Ethics Board approved this study. Hospital data accessed by GEMINI was approved by affiliated ethics boards.

Results

A total of 52 467 hospital admissions occurred between Jan. 23, 2020, and Dec. 31, 2020, and 11 852 of them were associated with at least 1 SARS-CoV-2 PCR test (regardless of result). In 444 of these 11 852 hospital admissions, we found at least 1 positive test, representing a frequency of 3.7% for confirmed SARS-CoV-2 infection. Among the 444 PCR-confirmed cases, 434 had an ICD-10 discharge diagnosis code U07.1 (Figure 1).

The sensitivity of code U07.1 was 97.8% (95% CI 95.9%–98.9%), the specificity was 99.5% (95% CI 99.3%–99.6%), the PPV was 88.2% (95% CI 85.0%–90.9%) and the NPV was 99.9% (99.8%–100.0%). Stratified analyses are presented in Tables 1 and 2. Operating characteristics were similar across sex, age groups and calendar periods (Table 1). When considering the timing of the first PCR test, the specificity of ICD code U07.1 was slightly lower when the test was done more than 48 hours after admission (Table 2). We also considered variability in the accuracy of coding by hospital (Table 3). The usage of U07.1 remained consistent over time (Figure 2).

Interpretation

Our results demonstrate high sensitivity, specificity and PPV estimates of ICD-10 code U07.1 in identifying laboratory-confirmed SARS-CoV-2 infection in hospital data. These operating characteristics were similar across sex, age groups and calendar periods. The sensitivity of the ICD code was higher when the PCR test was done within 24–48 hours of hospital admission. While our study found a high PPV for the ICD-10 code U07.1 in identifying laboratory-confirmed SARS-CoV-2 infection in hospital data, we acknowledge that this measurement can be influenced by the prevalence of

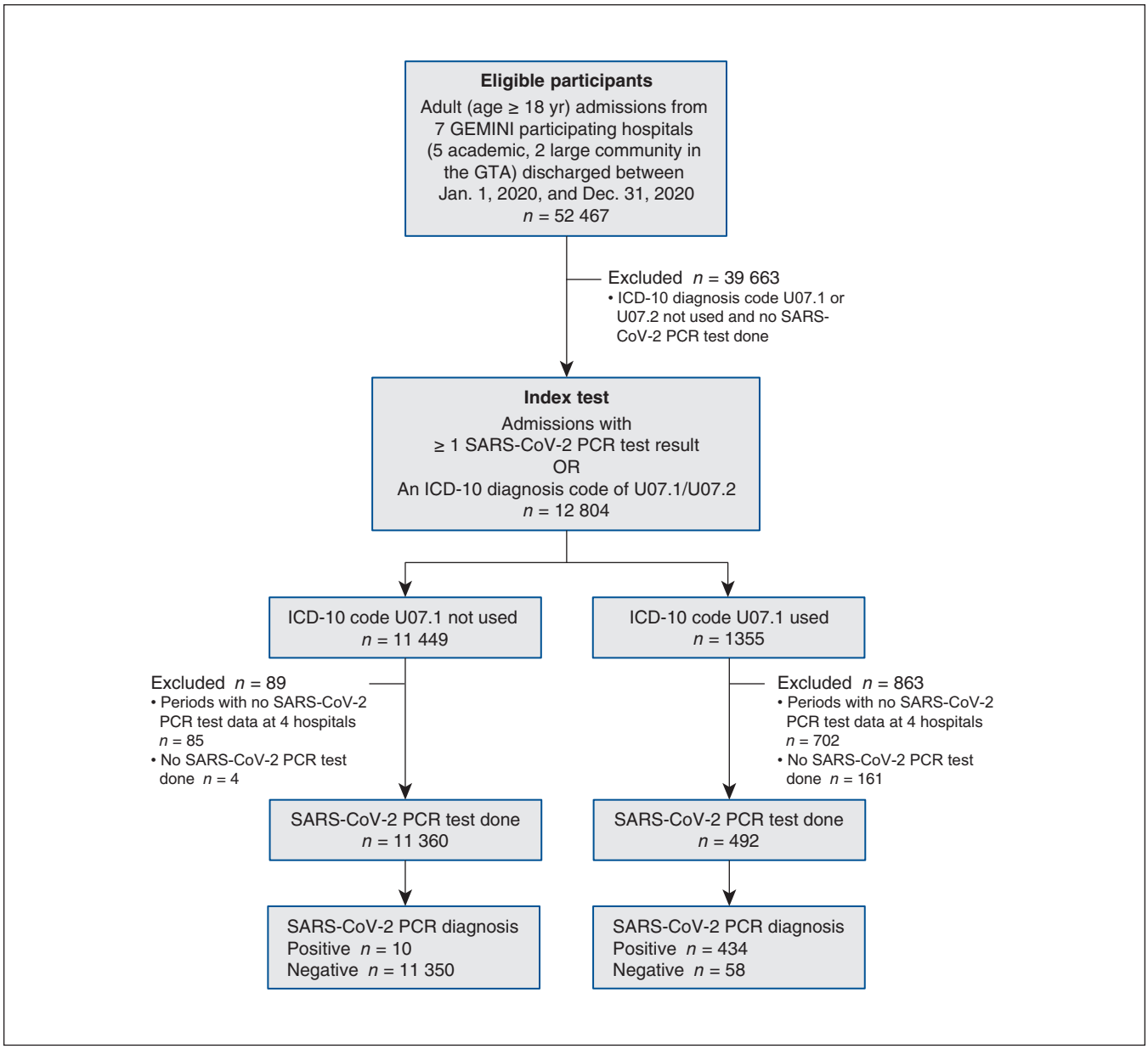

Figure 1: STARD (Standards for Reporting Diagnostic accuracy studies) flow diagram of the study cohort for the evaluation of the accuracy of the use of ICD-10 U07.1 diagnosis code. GEMINI = General Medicine Inpatient Initiative, GTA = Greater Toronto Area, ICD-10 = *International Statistical Classification of Diseases and Related Health Problems, 10th Revision*, PCR = polymerase chain reaction.

COVID-19 in the population being tested, as well as the criteria for PCR testing. For example, at the beginning of the pandemic, only more severe cases were tested as PCR tests were not widely available at that period. This selective testing approach may have led to an underrepresentation of mild or asymptomatic cases in the PCR-confirmed data. Moreover, it is important to consider the potential impact of the prolonged time interval between the onset of initial symptoms and hospital admission. This delay could result in declining viral load over time, potentially leading to PCR test results appearing as negative at the hospital admission. These factors contribute to the complexity of interpreting PCR results and underscore the need to consider the limitations of relying solely on testing data for assessing the true burden of SARS-CoV-2 infections.

Our results are consistent with those of studies in other jurisdictions evaluating the reliability of the ICD-10 code U07.1 in identifying SARS-CoV-2 infection in hospitalization data. In the only other Canadian study, Wu and colleagues assessed the validity of SARS-CoV-2 ICD codes in Alberta provincial health databases from emergency admissions and inpatients in 2 cohorts linked to administrative health records.[5] They found that the sensitivity of ICD-10 code U07.1 for inpatients with positive PCR tests was 94.2% (95%

Table 1: Operating characteristics of ICD code U07.1 (laboratory-confirmed COVID-19) within hospital diagnostic codes (PCR test as reference) stratified by sex, age, and urban or rural residence

Parameter	Estimate (95% CI), %
Sex	
Female	
Sensitivity	97.2 (93.6–99.1)
Specificity	99.5 (99.2–99.6)
Positive predictive value	86.1 (80.5–90.5)
Negative predictive value	99.9 (99.8–100)
Male	
Sensitivity	98.1 (95.7–99.4)
Specificity	99.5 (99.3–99.7)
Positive predictive value	89.7 (85.6–92.9)
Negative predictive value	99.9 (99.8–100)
Age	
< 50 yr	
Sensitivity	95.1 (87.8–98.6)
Specificity	99.5 (99.1–99.7)
Positive predictive value	84.6 (75.5–91.3)
Negative predictive value	99.8 (99.6–100)
50–75 yr	
Sensitivity	98.5 (95.6–99.7)
Specificity	99.5 (99.2–99.6)
Positive predictive value	87.1 (81.9–91.2)
Negative predictive value	99.9 (99.8–100)
> 75 yr	
Sensitivity	98.2 (94.8–99.6)
Specificity	99.6 (99.3–99.8)
Positive predictive value	91.5 (86.4–95.2)
Negative predictive value	99.9 (99.8–100)
Urban or rural residence	
Urban	
Sensitivity	97.6 (95.6–98.8)
Specificity	99.5 (99.3–99.6)
Positive predictive value	87.8 (84.5–90.7)
Negative predictive value	99.9 (99.8–100)
Rural	
Sensitivity	100 (39.8–100)
Specificity	99.7 (98.4–100)
Positive predictive value	80.0 (28.4–99.5)
Negative predictive value	100 (99.0–100)

Note: CI = confidence interval, ICD = *International Statistical Classification of Diseases and Related Health Problems*, PCR = polymerase chain reaction.

Table 2: Operating characteristics of ICD code U07.1 (laboratory-confirmed COVID-19) within hospital diagnostic codes (PCR test as reference) stratified by timing of first PCR test

Parameter	Estimate (95% CI), %
Timing of first PCR test	
Before admission	
Sensitivity	98.7 (96.2–99.7)
Specificity	99.9 (99.7–100)
Positive predictive value	97.8 (95.0–99.3)
Negative predictive value	99.9 (99.8–100)
Within 0–24 h of admission	
Sensitivity	95.8 (90.4–98.6)
Specificity	99.7 (99.6–99.9)
Positive predictive value	88.3 (81.4–93.3)
Negative predictive value	99.9 (99.8–100)
Within 24–48 h of admission	
Sensitivity	100 (63.1–100)
Specificity	99.0 (96.5–99.9)
Positive predictive value	80.0 (44.4–97.5)
Negative predictive value	100 (98.2–100)
Beyond 48 h of admission	
Sensitivity	97.8 (92.4–99.7)
Specificity	94.2 (92.1–95.9)
Positive predictive value	71.4 (62.7–79.1)
Negative predictive value	99.7 (98.8–100)
Calendar period of admission	
January–April 2020	
Sensitivity	99.4 (97.0–100)
Specificity	98.6 (97.9–99.2)
Positive predictive value	90.5 (85.5–94.2)
Negative predictive value	99.9 (99.6–100)
May–Aug 2020	
Sensitivity	96.2 (90.5–99.0)
Specificity	99.5 (99.2–99.7)
Positive predictive value	79.5 (71.5–86.2)
Negative predictive value	99.9 (99.8–100)
September–December 2020	
Sensitivity	96.8 (92.8–99.0)
Specificity	99.7 (99.6–99.9)
Positive predictive value	92.2 (87.0–95.8)
Negative predictive value	99.9 (99.8–100)

Note: CI = confidence interval, ICD = *International Statistical Classification of Diseases and Related Health Problems*, PCR = polymerase chain reaction.

Table 3: Accuracy of coding by hospital

Parameter	Estimate (95% CI), %
Confirmed cases (U07.1 only)	
Hospital A	
Sensitivity	98.5 (95.7–99.7)
Specificity	99.4 (99.2–99.6)
Positive predictive value	84.5 (79.3–88.9)
Negative predictive value	100 (99.9–100)
Hospital B	
Sensitivity	97.7 (94.1–99.4)
Specificity	99.8 (99.5–99.9)
Positive predictive value	95.4 (91.2–98.0)
Negative predictive value	99.9 (99.7–100)
Hospital C	
Sensitivity	95.9 (88.5–99.1)
Specificity	99.3 (98.8–99.6)
Positive predictive value	83.3 (73.6–90.6)
Negative predictive value	99.8 (99.5–100)

Note: CI = confidence interval.

CI 93.5%–94.8%), and the PPV was 94.5% (95% CI 93.8%–95.2%).[5]

Kadri and colleagues examined the reliability of ICD-10 code U07.1 in American administrative hospitalization data early in the pandemic.[11] Using a positive PCR test as the gold standard, they estimated the sensitivity of U07.1 at 98.0% (95% CI 97.6%–98.4%) specificity at 99.0% (95% CI 98.9%–99.1%) and PPV at 91.5% (95% CI 90.8%–92.3%).[11] They concluded that hospitals accurately code SARS-CoV-2 diagnoses, though they advocated for continuing to assess the reliability of this code over time.[11] Subsequent studies found similar results.[3,12,13] Lynch and colleagues in the United States reviewed Veterans Affairs medical records containing the ICD diagnostic code U07.1 and calculated PPV, with PCR testing as the reference standard.[4] They also found high PPVs in patients admitted to hospital (93.8%, 95% CI 91.8%–95.6%).[4] Bhatt and colleagues evaluated hospital discharge diagnoses between Apr. 1, 2020, and July 31, 2020, in the Mass General Brigham health system (which includes Massachusetts General Hospital, Brigham and Women's Hospital, and other allied hospitals across Massachusetts).[14] Compared with all other studies, they found a much lower sensitivity (49.2%, 95% CI 47.1%–51.3%) for the hospital ICD code U07.1 compared with PCR test results; they estimated high specificity (99.4%, 95% CI 99.3%–99.5%) and a PPV of 90.0% (95% CI 88.2%–91.6%).[14] They attributed the lower sensitivity to scribing

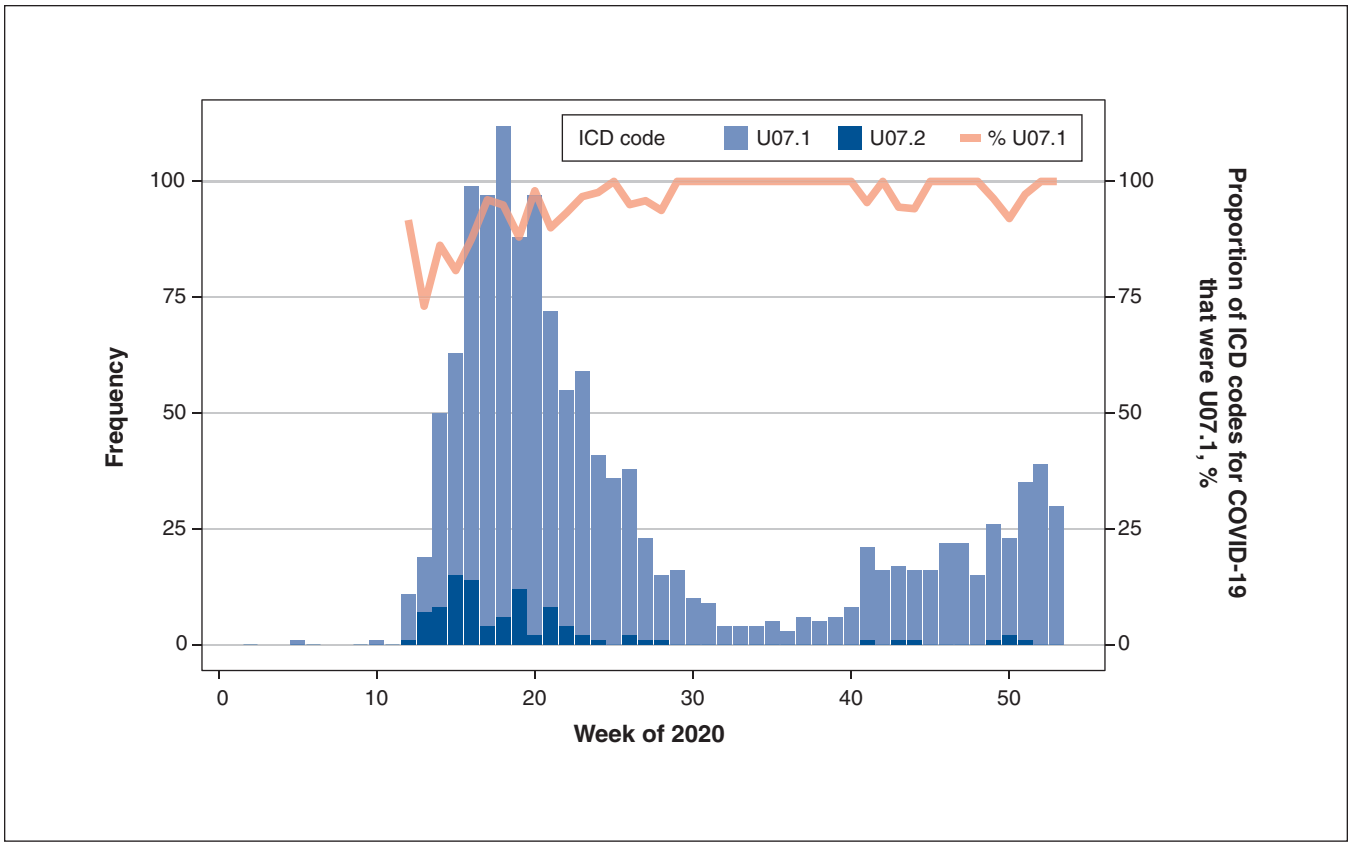

Figure 2: Plot of U07.1 and U07.2 ratio (i.e., laboratory-confirmed and clinical diagnosis) over time. Note: ICD = *International Statistical Classification of Diseases and Related Health Problems*.

delays at discharge, changes to testing criteria, and interpretation differences when looking at test results and symptom presentation.[14] Bodilsen and colleagues also confirmed a high PPV for SARS-CoV-2 codes in Danish hospital data.[15]

One of the motivations for using Canadian data for our study is that, in this country, individuals have publicly funded health care, so administrative records hold discharge diagnoses for all hospital admissions. Given that, to our knowledge, only one other Canadian study has been published on this topic,[5] and it is well known that use of ICD codes (and their validation) can vary greatly over jurisdictions, our study is important. Moreover, since the sensitivity and specificity of case definitions from administrative data can vary greatly over demographic and other variables, our stratified analyses offer some unique information.

Limitations

Our study has some potential limitations. First, PCR tests performed outside of Ontario hospitals were unavailable, which may have caused some individuals tested for SARS-CoV-2 in outpatient settings to be identified as false positives (i.e., with a U07.1 diagnosis but without a positive test). However, this limitation is common in many published studies on this topic. Additionally, our study uses data from the first year of the pandemic, and as others have suggested, repeat analyses in years to come may yield further insights.[16] Testing policies for SARS-CoV-2 varied across hospitals and fluctuated over time (Table 2). Still, generally, all patients with typical or atypical symptoms associated with COVID-19 were tested, and, in some settings, all admitted patients were tested. Our findings are generally consistent across periods, and policy changes may explain minor fluctuation.

Conclusion

Overall, our findings confirm the validity of the ICD-10 code U07.1. We found consistent results across sex, age groups and calendar periods. This is reassuring for research and surveillance activities relying on administrative hospitalization data to identify SARS-CoV-2 infections.

References

1. Important information for CCRS, HCRS, IRRS, OMHRS and NRS. Ottawa: Canadian Institute for Health Information; 2020. Available: https://www.cihi.ca/en/important-information-for-ccrs-hcrs-irrs-omhrs-and-nrs (accessed 2022 Sept. 22).
2. Update: Capturing COVID-19-related diagnoses in the NRS. Ottawa: Canadian Institute for Health Information; 2020. Available: https://www.cihi.ca/en/update-capturing-covid-19-related-diagnoses-in-the-nrs (accessed 2023 July 1).
3. Moll K, Hobbi S, Zhou CK, et al. Assessment of performance characteristics of COVID-19 ICD-10-CM diagnosis code U07.1 using SARS-CoV-2 nucleic acid amplification test results. *PLoS One* 2022;17:e0273196. doi: 10.1371/journal.pone.0273196.
4. Lynch KE, Viernes B, Gatsby E, et al. Positive predictive value of COVID-19 ICD-10 diagnosis codes across calendar time and clinical setting. *Clin Epidemiol* 2021;13:1011-8.
5. Wu G, D'Souza AG, Quan H, et al. Validity of ICD-10 codes for COVID-19 patients with hospital admissions or ED visits in Canada: a retrospective cohort study. *BMJ Open* 2022;12:e057838.
6. Verma AA, Guo Y, Kwan JL, et al. Patient characteristics, resource use and outcomes associated with general internal medicine hospital care: the General Medicine Inpatient Initiative (GEMINI) retrospective cohort study. *CMAJ Open* 2017;5:E842-9.
7. Verma AA, Pasricha SV, Jung HY, et al. Assessing the quality of clinical and administrative data extracted from hospitals: the General Medicine Inpatient Initiative (GEMINI) experience. *J Am Med Inform Assoc* 2021;28:578-87.
8. Canadian coding standards for version 2018 ICD-10-CA and CCI. Ottawa: Canadian Institute for Health Information; 2018. Available: https://secure.cihi.ca/free_products/CodingStandards_v2018_EN.pdf (accessed 2023 July 1).
9. *Postal CodeOM Conversion File Plus (PCCF+) version 7B, reference guide*. Ottawa: Statistics Canada, 2019.
10. R version 4.1.2. The R Project for Statistical Computing. Indianapolis (IN): The R Foundation. Available: https://www.R-project.org/ (accessed 2023 June 22).
11. Kadri SS, Gundrum J, Warner S, et al. Uptake and accuracy of the diagnosis code for COVID-19 among US hospitalizations. *JAMA* 2020;324:2553-4.
12. Kluberg SA, Hou L, Dutcher SK, et al. Validation of diagnosis codes to identify hospitalized COVID-19 patients in health care claims data. *Pharmacoepidemiol Drug Saf* 2022;31:476-80.
13. Rao S, Bozio C, Butterfield K, et al. Accuracy of COVID-19–like illness diagnoses in electronic health record data: retrospective cohort study. *JMIR Form Res* 2023;7:e39231.
14. 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;36:2532-5.
15. Bodilsen J, Leth S, Nielsen SL, et al. Positive predictive value of ICD-10 diagnosis codes for COVID-19. *Clin Epidemiol* 2021;13:367-72.
16. Arevalo-Rodriguez I, Buitrago-Garcia D, Simancas-Racines D, et al. False-negative results of initial RT-PCR assays for COVID-19: a systematic review. *PLoS One* 2020;15:e0242958. doi: 10.1371/journal.pone.0242958.

Affiliations: Faculty of Medicine (Moura, Bernatsky), McGill University; Research Institute of the McGill University Health Centre (Neville, Bernatsky), Montréal, Que.; Li Ka Shing Knowledge Institute, St. Michael's Hospital (Liao, Wen, Razak, Roberts, Verma), Unity Health Toronto; Department of Medicine (Razak, Verma) and Institute of Health Policy, Management and Evaluation (Razak, Roberts, Verma), University of Toronto, Toronto, Ont.

Contributors: Cristiano Moura and Sasha Bernatsky contributed to the study design. Fangming Liao, Bijun Wen, Fahad Razak, Surain Roberts and Amol Verma contributed to data collection and analysis. All authors contributed to the interpretation of results, and preparation and editing of the manuscript. All authors gave final approval of the version to be published and agreed to be accountable for all aspects of the work.

Funding: This work was supported by the Canadian Institutes of Health Research [DMC-166262].

Data sharing: The authors are unable to provide unlimited open access to GEMINI data because of data sharing agreements and research ethics board protocols with participating hospitals. However, researchers can request access to GEMINI data through an established process approved by GEMINI's institutional research ethics boards. Please see full details at https://www.geminimedicine.ca/access-data.

Acknowledgements: GEMINI acknowledgement: The authors acknowledge the individuals and organizations that have made the data available for this research. The development of the GEMINI data platform has been supported with funding from the Canadian Cancer Society, the Canadian Frailty Network, the Canadian Institutes of Health Research, the Canadian Medical Protective Association, Green Shield Canada Foundation, the Natural Sciences and Engineering Research Council of Canada, Ontario Health, the St. Michael's Hospital Association Innovation Fund, the University of Toronto Department of Medicine, and in-kind support from partner hospitals and Vector Institute.

Supplemental information: For reviewer comments and the original submission of this manuscript, please see www.cmajopen.ca/content/11/5/E982/suppl/DC1.