BMJ Open

Validity of ICD-10 codes for COVID-19 patients with hospital admissions or ED visits in Canada: a retrospective cohort study

Guosong Wu ,[1] Adam G D'Souza,[1,2] Hude Quan,[1] Danielle A Southern,[1] Erik Youngson,[2] Tyler Williamson,[1] Cathy Eastwood ,[1] Yuan Xu[1,3]

To cite: 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. doi:10.1136/bmjopen-2021-057838

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-057838).

Received 30 September 2021
Accepted 20 December 2021

[1]Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada
[2]Alberta Health Services, Calgary, Alberta, Canada
[3]Department of Oncology and Surgery, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada

Correspondence to
Dr Guosong Wu;
guosong.wu@ucalgary.ca

ABSTRACT

Objective To evaluate the validity of COVID-19 International Classification of Diseases, 10th Revision (ICD-10) codes and their combinations.

Design Retrospective cohort study.

Setting Acute care hospitals and emergency departments (EDs) in Alberta, Canada.

Participants Patients who were admitted to hospital or presented to an ED in Alberta, as captured by local administrative databases between 1 March 2020 and 28 February 2021, who had a positive COVID-19 test and/or a COVID-19-related ICD-10 code.

Main outcome measures The sensitivity, positive predictive value (PPV) and 95% CIs for ICD-10 codes were computed. Stratified analysis on age group, sex, symptomatic status, mechanical ventilation, hospital type, patient intensive care unit (ICU) admission, discharge status and season of pandemic were conducted.

Results Two overlapping subsets of the study population were considered: those who had a positive COVID-19 test (cohort A, for estimating sensitivity) and those who had a COVID-19-related ICD-10 code (cohort B, for estimating PPV). Cohort A included 17 979 ED patients and 6477 inpatients while cohort B included 33 675 ED patients and 18 746 inpatients. Of inpatients, 9.5% in cohort A and 8.1% in cohort B received mechanical ventilation. Over 13% of inpatients were admitted to ICU. The length of hospital stay was 6 days (IQR: 3–14) for cohort A and 8 days (IQR: 3–19) for cohort B. In-hospital mortality was 15.9% and 38.8% for cohort A and B, respectively. The sensitivity for ICD-10 code U07.1 (COVID-19, virus identified) was 82.5% (81.8%–83.2%) with a PPV of 93.1% (92.6%–93.6%). The combination of U07.1 and U07.3 (multisystem inflammatory syndrome associated with COVID-19) had a sensitivity of 82.5% (81.9%–83.2%) and PPV of 92.9% (92.4%–93.4%).

Conclusions In Alberta, ICD-10 COVID-19 codes (U07.1 and U07.3) were coded well with high validity. This indicates administrative data can be used for COVID-19 research and pandemic management purposes.

Strengths and limitations of this study

- ► This is the first endeavour to explore the validity of COVID-19-related International Classification of Diseases, 10th Revision (ICD-10) codes using both outpatients and hospitalised patients.
- ► With a large population-based retrospective cohort study, the epidemiology, susceptibility and outcomes of COVID-19 were summarised, the sensitivity and positive predictive valued of ICD-10 codes were computed with data collected over an entire pandemic year.
- ► Validity of ICD-10 codes for COVID-19 and their combinations was computed and stratified analysis presented by patient demographic and clinical characteristics.
- ► While the study presents the sensitivity and positive predictive value, the specificity and negative predictive value could not be determined because the data cannot be used to reliably estimate the true negatives.
- ► The extent to which the research findings can be generalised to other countries or healthcare settings is unknown.

INTRODUCTION

Since the declaration of a pandemic by the WHO, SARS-CoV-2 has caused 195.9 million infections and caused over 4.2 million deaths globally.[1] An enormous number of research projects has been conducted to better understand the disease and its impact.[2] For example, there are real-world evidence studies pertaining to the long-term effect on health of survivors,[3] large-scale epidemiological studies to explore the natural history of disease outcomes,[4] and population-based health services research and policy studies to explore the optimal coping strategies for future outbreaks.[5 6] However, case identification of COVID-19 is the first critical step for all these initiatives.

In a quick response to the pandemic, WHO activated two emergency International Classification of Diseases, 10th Revision (ICD-10) codes for COVID-19 in February 2020, U07.1 for confirmed cases and U07.2 for suspected

or probable cases (clinical or epidemiological diagnosis).[7] A set of additional codes was defined later on to capture COVID-19-related information.[8] To date, there is limited information on the performance of ICD-10 codes in identifying COVID-19 patients who were admitted to hospitals or visited emergency departments (EDs). Estimates of the validity of U07.1 among hospitalised patients have varied (range 49%–98%) across countries and over time.[9–11] As the pandemic continues to evolve, it is important to assess the validity of ICD-10 codes using large population-based data from the past pandemic year and provide accurate algorithms to identify COVID-19 cases.

This study sought to evaluate the validity of ICD-10 codes in identifying individuals who experienced COVID-19 through population-based administrative databases with laboratory test results as reference standard.

METHODS

We conducted a diagnostic coding accuracy study on a consecutive cohort of COVID-19 patients in Alberta, Canada.

Study cohort

This retrospective cohort study included all patients who were diagnosed with COVID-19 and had an ED visit or were admitted to a hospital in Alberta, Canada between 1 March 2020 and 28 February 2021. Only first records were analysed if a patient had multiple encounters in hospitalisation or ED visits.

A patient was defined as a COVID-19 case if they had an ED visit or hospitalisation that occurred between 1 day prior to, and up to 7 days after a positive SARS-CoV-2 PCR test recorded in a laboratory database. Different cut-offs for earliest and latest dates of encounters relative to the positive test date were tested, with no significant impact to any of the reported sensitivity or positive predictive value (PPV) results.

The validity of the ICD-10 codes and their combinations was calculated from the following two cohorts. Cohort A contained all positive COVID-19 cases, linked to administrative databases to calculate sensitivity. Cohort B included all patients who were assigned one of the COVID-19-related ICD codes in administrative databases, linked back to the laboratory database to determine if a positive PCR test existed, to calculate PPV.

Data sources

The data were derived from three Alberta provincial databases that cover the Alberta population: (1) Discharge Abstract Database (DAD), which contains demographic, administrative and clinical data for hospitalised patients including up to 25 ICD-10 diagnosis codes per record; (2) National Ambulatory Care Reporting System (NACRS), which captures data of hospital-based ambulatory care outpatient clinics, day surgery and ED visits and (3) Public Health Laboratory (ProvLab) database, which captures SARS-CoV-2 laboratory PCR test results and dates and was used as the reference standard. The patient personal health number, sex, and date of birth were used to conduct the data linkage. Deidentified data were received from Alberta Health Services and analysed within a secure computing environment at the University of Calgary.

ICD-10 codes of COVID-19

The Canadian Institute for Health Information (CIHI) updates COVID-19 coding directions when new codes are released by WHO. All codes[8] that were used by CIHI during the pandemic were included in this study. This includes U07.1 (COVID-19, virus identified), U07.3 (Multisystem inflammatory syndrome associated with COVID-19), O98.5 (COVID-19 in pregnancy), Z03.8 (Observation for other suspected diseases), Z11.5 (Encounter for screening for other viral diseases), Z51.5 (COVID-19 in palliative care) and Z71.1 (Person with feared complaint in whom no diagnosis is made). We also assessed the validity of two combinations of codes, set 1: U07.1 and U07.3, and set 2: U07.1, U07.3, O98.5, Z03.8, Z11.5, Z51.5 and Z71.1. ICD-10 code U07.2 (virus not identified) is assigned when the patient is diagnosed, clinically or epidemiologically, with an acute infection with the COVID-19 virus, but the COVID-19 PCR lab test results are inconclusive or not available, or no test was performed.[8] Since PCR lab test was used as the gold standard, it was not suitable for assessing the validity of this code. Therefore, code U07.2 was excluded from this study.

Statistical analysis

Descriptive statistics were used to report characteristics of the study cohorts. Charlson Comorbidity Index was derived from DAD and NACRS.[12] Sensitivity and PPV were calculated through comparing ICD-10 codes in administrative data against the reference standard from ProvLab, and 95% binomial proportion 95% CIs were computed using the Wilson method. We estimated the overall performance of ICD-10 codes, and subgroup performance stratified by patient characteristics (eg, age group, sex, mechanical ventilation), hospital type, outcome variables (intensive care unit, ICU admission, discharge status) and seasons of pandemic for both study cohorts. All statistical analyses were performed using Python V.3 and STATA 17 software (StataCorp. 2021. Stata Statistical Software: Release 17: StataCorp.).

Patient and public involvement

Study participants and other members of the public were not involved in the design, or conduct, or reporting, or dissemination plans of the research.

RESULTS

A total of 17 979 ED patients and 6477 inpatients were included in cohort A, and 33 675 ED patients and 18 746 inpatients were included in cohort B (table 1). Overall, compared with the hospitalised patients, ED patients

Table 1 Baseline patient characteristics

Characteristic*	Cohort A		Cohort B	
	ED (N=17 979)	Inpatient (N=6477)	ED (N=33 675)	Inpatient (N=18 746)
Age in years median (IQR)	47 (31–64)	64 (46–79)	43 (26–63)	73 (58–84)
≤18	1301 (7.24)	173 (2.67)	5034 (14.95)	581 (3.10)
19–40	5954 (33.12)	1086 (16.77)	10 420 (30.94)	1596 (8.51)
41–60	5429 (30.20)	1622 (25.04)	8600 (25.54)	3059 (16.32)
60–80	3610 (20.08)	2174 (33.56)	6637 (19.71)	7291 (38.89)
>80	1685 (9.37)	1422 (21.95)	2984 (8.86)	6219 (33.18)
Female	9009 (50.11)	2986 (46.11)	17 411 (51.7)	9110 (48.60)
Charlson Index, median (IQR)	0 (0–2)	2 (0–4)	0 (0–2)	4 (1–6)
Myocardial infarction	1014 (5.64)	759 (11.72)	1697 (5.04)	2757 (14.71)
Chronic heart failure	1186 (6.60)	1006 (15.53)	2034 (6.04)	4230 (22.56)
Peripheral vascular disease	548 (3.05)	437 (6.75)	1007 (2.99)	1831 (9.77)
Cerebrovascular disease	1082 (6.02)	815 (12.58)	1998 (5.93)	3711 (19.80)
Dementia	828 (4.61)	763 (11.78)	1239 (3.68)	2843 (15.17)
Chronic obstructive pulmonary disease	3519 (19.57)	1737 (26.82)	6280 (18.65)	5370 (28.65)
Rheumatoid arthritis	334 (1.86)	206 (3.18)	564 (1.67)	678 (3.62)
Peptic ulcer disease	732 (4.07)	417 (6.44)	1235 (3.67)	1489 (7.94)
Liver disease	951 (5.29)	644 (9.94)	1719 (5.11)	2462 (13.13)
Diabetes	5587 (31.07)	3829 (59.11)	8227 (24.43)	10 159 (54.19)
Paraplegia and hemiplegia	315 (1.75)	246 (3.80)	521 (1.55)	1102 (5.88)
Renal disease	793 (4.41)	675 (10.42)	1299 (3.86)	2673 (14.26)
Cancer	1105 (6.15)	789 (12.18)	3588 (10.65)	6931 (36.97)
Metastatic carcinoma	280 (1.56)	225 (3.47)	1934 (5.69)	4449 (23.73)
AIDS/HIV	43 (0.24)	31 (0.48)	86 (0.26)	61 (0.33)
Any comorbidity	8278 (46.04)	4644 (71.70)	14 478 (42.99)	15 988 (85.29)
Symptomatic at test				
Yes	8031 (44.67)	2343 (36.17)	5124 (15.22)	1977 (10.55)
No	4288 (23.85)	1434 (22.14)	2207 (6.55)	1135 (6.05)
Unknown	1964 (10.92)	1014 (15.66)	1361 (4.04)	872 (4.65)
Missing	3696 (20.56)	1686 (26.03)	24 983 (74.19)	14 762 (78.75)
Mechanical ventilation	NA	618 (9.54)	NA	1515 (8.08)
Days between positive lab test to admission or ED visit, median (IQR)	0 (–1 to 7)	1 (0, 7)	1 (–1 to 6)	1 (0, 6)
ICU admission	NA	970 (14.98)	NA	2448 (13.06)
Length of Stay (days), median (IQR)	NA	6 (3–14)	NA	8 (3–19)
Death at discharge	34 (0.19)	1029 (15.89)	161 (0.48)	7276 (38.81)
Location of admission				
Suburban/rural	5453 (30.33)	1124 (17.35)	11 226 (33.34)	3822 (20.39)
Large urban	6285 (34.96)	3186 (49.19)	10 580 (31.12)	7244 (38.64)
Others	6241 (34.71)	2167 (33.46)	11 849 (35.18)	7680 (40.97)

*Data presented as number (percentage) of patients unless otherwise indicated.
ED, emergency department; ICU, intensive care unit; NA, not applicable.

were more likely to be younger (cohort A: median age 47 vs 64, cohort B: median age 43 vs 73) females (cohort A: 50.1% vs 46.1%, cohort B: 51.7% vs 48.6%). Hospitalised patients in cohort B were more likely to have cancers (36.97% vs 12.18%), particularly metastatic carcinoma (23.73% vs 3.47%), compared with cohort A. Of hospitalised patients, 9.5% in cohort A and 8.1% in cohort B received mechanical ventilation. There were about half as many flagged asymptomatic cases at the time of testing as flagged symptomatic cases in both cohorts. Of the hospitalised patients, 15.0% and 13.1% patients were admitted to ICU in cohort A and B, respectively. The

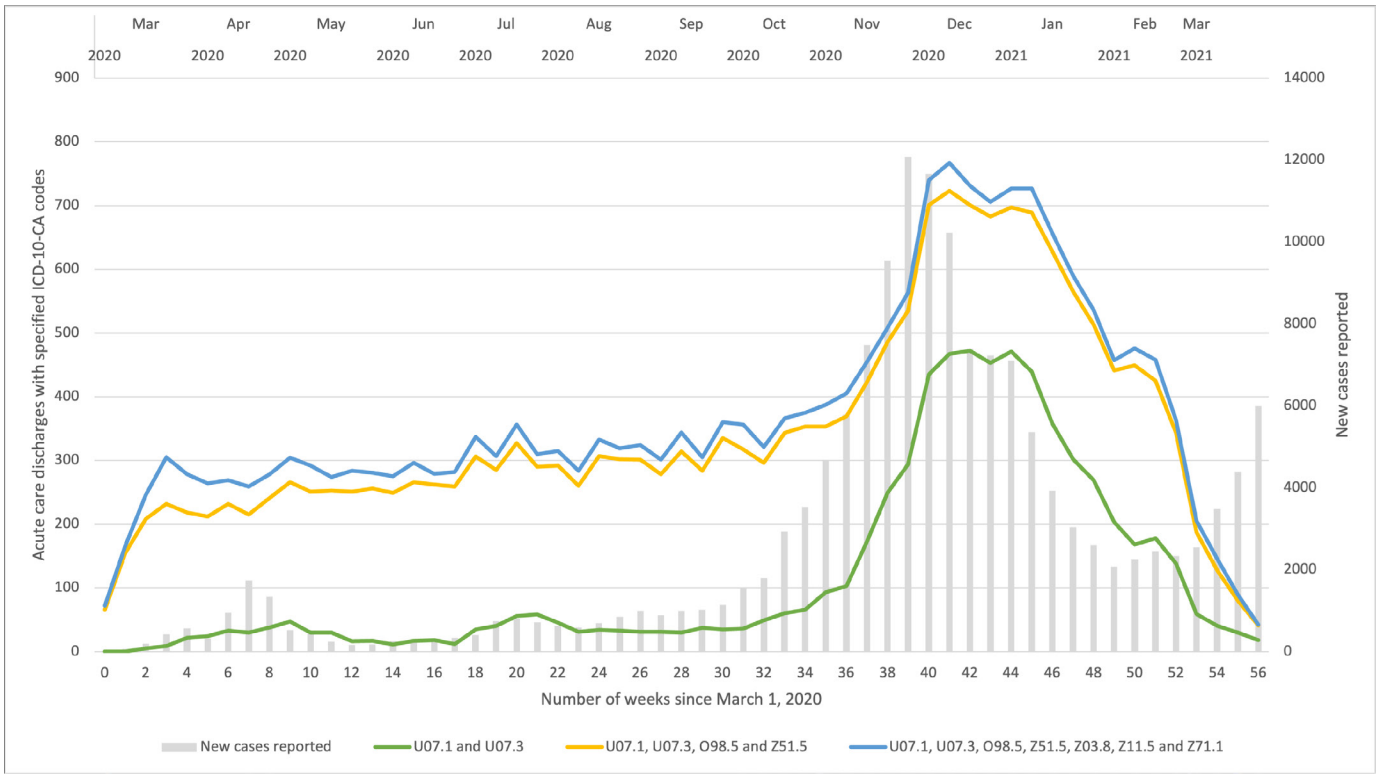

Figure 1 COVID-19-related ICD-10 code counts among inpatients and ED visits (left line chart) and new cases reported (right bar chart) between 1 March 2020 and 28 February 2021. ED, emergency department; ICD-10, International Classification of Diseases, 10th Revision.

length of hospital stay was 6 days (IQR: 3–14) for Cohort A and 8 days (IQR: 3–19) for cohort B. In-hospital mortality was 15.9% and 38.8% for cohort A and B, respectively. The week-by-week COVID-19-related ICD-10 code counts among inpatients and ED visits ranged from 72 counts in March to 767 in December 2020 (figure 1).

For code U07.1, the sensitivity was 82.5% (95% CI 81.8% to 83.2%) and PPV was 93.1% (95% CI 92.6% to 93.6%) (table 2). Compared with ED visits, inpatients had higher sensitivity (94.2% vs 81.3%) and similar PPV (94.5% vs 93.3%). The combination of codes U07.1 and U07.3 for entire cohort had a sensitivity of 82.5% (95% CI 81.9% to 83.2%) and PPV of 92.9% (95% CI 92.4% to 93.4%). The combination of all related codes had a sensitivity of 84.4% (95% CI 83.8% to 85.1%) but PPV of 23.1% (95% CI 22.7% to 23.5%).

Stratified analysis of code U07.1 over the entire cohort shows higher sensitivity and PPV for patients aged 80 and above, patients who were admitted to ICU, ventilated patients, and inpatient survivors relative to encounters (table 3). The validity of U07.1 varied by season, with higher PPV in summer (95.8%, 95% CI 94.2% to 96.9%), and higher sensitivity in spring (86.2%, 95% CI 83.8%

Table 2 Performance characteristics of ICD-10 among inpatients and ED visits

	ED visits		Inpatients		Entire cohort	
ICD-10 codes	Sensitivity (95% CI)	PPV (95% CI)	Sensitivity (95% CI)	PPV (95% CI)	Sensitivity (95% CI)	PPV (95% CI)
U07.1	81.27% (80.53% to 81.99%)	93.34% (92.83% to 93.83%)	94.19% (93.49% to 94.82%)	94.51% (93.78% to 95.17%)	82.48% (81.79% to 83.15%)	93.09% (92.59% to 93.56%)
U07.1 and U07.3	81.31% (80.57% to 82.02%)	93.24% (92.72% to 93.73%)	94.42% (93.74% to 95.04%)	94.19% (93.43% to 94.86%)	82.54% (81.85% to 83.21%)	92.91% (92.40% to 93.39%)
U07.1, U07.3, O98.5, Z03.8, Z11.5, Z51.5, Z71.1	83.30% (82.59% to 83.99%)	29.22% (28.72% to 29.72%)	95.83% (95.22% to 96.36%)	22.91% (22.29% to 23.55%)	84.42% (83.76% to 85.06%)	23.09% (22.69% to 23.49%)

ED, emergency department; ICD-10, International Classification of Diseases, 10th Revision; PPV, positive predictive value.

Table 3 Performance of ICD-10 (U07.1) among patient subgroups tested for confirmed COVID-19

	ED visits		Inpatients		Entire cohort	
Characteristic	**Sensitivity (95% CI)**	**PPV (95% CI)**	**Sensitivity (95% CI)**	**PPV (95% CI)**	**Sensitivity (95% CI)**	**PPV (95% CI)**
Age, years						
≤18	76.55% (73.26% to 79.55%)	93.55% (91.20% to 95.30%)	82.18% (73.58% to 88.42%)	86.36% (77.66% to 92.02%)	77.07% (73.87% to 79.99%)	92.63% (90.25% to 94.46%)
19–40	79.16% (77.72% to 80.53%)	90.80% (89.67% to 91.82%)	89.49% (86.93% to 91.60%)	89.11% (86.47% to 91.29%)	79.07% (77.66% to 80.42%)	90.28% (89.17% to 91.29%)
41–60	82.87% (81.59% to 84.08%)	93.95% (93.06% to 94.74%)	95.18% (93.82% to 96.25%)	94.74% (93.28% to 95.90%)	83.92% (82.69% to 85.07%)	93.92% (93.05% to 94.69%)
61–80	82.27% (80.73% to 83.72%)	94.92% (93.92% to 95.76%)	95.17% (94.04% to 96.09%)	96.09% (94.97% to 96.97%)	83.89% (82.49% to 85.21%)	94.57% (93.59% to 95.40%)
>80	83.06% (80.67% to 85.21%)	95.21% (93.61% to 96.43%)	95.56% (94.18% to 96.63%)	96.26% (94.74% to 97.35%)	86.94% (85.04% to 88.63%)	95.20% (93.73% to 96.34%)
Sex						
Male	81.22% (80.18% to 82.23%)	93.14% (92.39% to 93.81%)	94.77% (93.85% to 95.55%)	94.56% (93.57% to 95.41%)	82.76% (81.78% to 83.69%	92.90% (92.18% to 93.56%
Female	81.31% (80.25% to 82.33%)	93.56% (92.82% to 94.23%)	93.48% (92.34% to 94.45%)	94.40% (93.23% to 95.37%)	82.19% (81.19% to 83.15%	93.27% (92.55% to 93.93%
Ventilation						
Yes	NA	NA	96.76% (95.00% to 97.92%)	94.40% (92.17% to 96.03%)	96.76% (95.00% to 97.92%	94.40% (92.17% to 96.03%
No	NA	NA	93.81% (93.05% to 94.50%)	94.33% (93.53% to 95.04%)	82.29% (81.59% to 82.97%	93.13% (92.63% to 93.61%
ICU admission						
Yes	NA	NA	97.03% (95.71% to 97.95%)	94.46% (92.71% to 95.81%)	97.03% (95.71% to 97.95%	94.46% (92.71% to 95.81%
No	NA	NA	93.48% (92.67% to 94.20%)	94.30% (93.47% to 95.03%)	93.48% (92.67% to 94.20%	94.30% (93.47% to 95.03%
Patient discharge status						
Alive	81.30% (80.56% to 82.02%)	93.37% (92.85% to 93.85%)	93.67% (92.86% to 94.40%)	94.15% (93.31% to 94.89%)	82.09% (81.38% to 82.77%)	93.07% (92.56% to 93.55%)
Dead	69.23% (50.01% to 83.50%)	85.00% (63.96% to 94.76%)	96.19% (94.79% to 97.22%)	95.71% (94.04% to 96.93%)	95.48% (94.01% to 96.61%)	95.44% (93.76% to 96.69%)
Academic hospital						
Yes	80.06% (78.07% to 81.91%)	93.49% (92.07% to 94.67%)	95.56% (94.08% to 96.69%)	91.99% (90.00% to 93.61%)	83.28% (81.52% to 84.90%)	92.03% (90.63% to 93.23%)
No	81.30% (80.56% to 82.02%)	93.14% (92.57% to 93.66%)	93.75% (92.95% to 94.47%)	94.73% (93.92% to 95.44%)	81.69% (80.93% to 82.42%)	93.02% (92.47% to 93.53%)
Season change						
March–May	84.60% (82.01% to 86.88%)	92.54% (90.40% to 94.24%)	95.83% (93.00% to 97.55%)	92.42% (88.59% to 95.04%)	86.22% (83.82% to 88.32%)	92.38% (90.31% to 94.03%)
June–August	75.25% (72.50% to 77.80%)	96.09% (94.53% to 97.22%)	94.28% (91.41% to 96.23%)	96.39% (93.66% to 97.97%)	76.73% (74.14% to 79.14%)	95.75% (94.23% to 96.89%)
September–November	78.62% (77.16% to 80.00%)	95.33% (94.46% to 96.06%)	92.50% (90.99% to 93.78%)	95.55% (94.12% to 96.65%)	80.20% (78.87% to 81.47%)	94.82% (93.95% to 95.57%)
December–February	83.15% (82.18% to 84.07%)	91.66% (90.90% to 92.36%)	94.44% (93.52% to 95.23%)	93.97% (92.98% to 94.82%)	84.05% (83.13% to 84.92%)	91.54% (90.81% to 92.22%)
Symptomatic at test						
Yes	80.10% (78.93% to 81.22%)	100.00% (99.90% to 100.00%)	95.40% (94.26% to 96.32%)	100.00% (99.74% to 100.00%)	80.95% (79.82% to 82.03%)	100.00% (99.90% to 100.00%)
No	80.12% (78.03% to 82.06%)	100.00% (99.68% to 100.00%)	92.52 (90.39% to 94.21%)	100.00 (99.36% to 100.00%)	81.57% (79.64% to 83.36%)	100.00% (99.97% to 100.00%)
Unknown	75.02% (72.81% to 77.10%)	100.00% (99.68% to 100.00%)	91.68% (89.68% to 93.32%)	100.00% (99.40% to 100.00%)	77.80% (75.83% to 79.65%)	100.00% (99.72% to 100.00%)
Missing	86.23% (85.01% to 87.37%)	80.30% (78.95% to 81.58%)	94.52% (93.28% to 95.54%)	83.65% (81.68% to 85.44%)	86.97% (85.84% to 88.03%)	79.72% (78.42% to 80.96%)

ED, emergency department; ICD-10, International Classification of Diseases, 10th Revision; ICU, intensive care unit; PPV, positive predictive value.

to 88.3%). The sensitivity and PPV were similar between symptomatic and asymptomatic patients.

DISCUSSION

Our study demonstrated that the ICD-10 code U07.1 for SARS-CoV-2 disease had high sensitivity and PPV. Adding other COVID-19-related codes increased the sensitivity but decreased the PPV. The sensitivity and PPV varied between outpatient and inpatient cohorts, as well as by patient characteristics.

Our findings indicated that ICD-10 code U07.1 accurately identified COVID-19 cases within the administrative database in Alberta, Canada. Recent studies from other care settings or countries evaluated the validity of code U07.1 with 3–5 months of observation.[9 10 13] Our study retrospectively analysed the code validity for the past pandemic year and found that validity of administrative data in recording COVID-19 varied by seasons, as well as by patient characteristics such as age, admission to ICU, and discharge status (alive or dead).

The study cohorts A and B are similarly distributed in most aspects (proportions of inpatients, ages of ED vs inpatients and many of the comorbidities), but stark differences were observed in the frequencies of certain severe health conditions (eg, cohort B were more likely to have cerebrovascular disease and cancers). This may be because using ICD codes to identify COVID-19 patients in cohort B might be more likely to capture patients with mixed primary diagnoses, whereas using positive COVID-19 PCR test results to define subsequent in-hospital COVID-19 patients in cohort A was more likely to capture COVID-19 patients who were hospitalised primarily due to their COVID-19 diagnosis.

To the best of our knowledge, this is the first endeavour to explore the validity of COVID-19-related ICD-10 codes using both outpatients and hospitalised patients, based on our review of the literature.[9–11] We analysed a large population-based database and provided robust evidence for the validity of the ICD-10 codes. Combinations of different sets of COVID-19-related ICD codes could slightly improve the sensitivity but doing so would, however, compromise the PPV. The observed sensitivity and PPV were higher in the hospitalised patient cohort compared with the ED visitors. Depending on their investigative purpose, researchers need to choose the best method for COVID-19 case identification with administrative databases.

The sensitivity and PPV of U07.1 were observed to be higher in patients aged 80 and above as well as in patients with severe health conditions or even death. A similar pattern was reported by Bhatt *et al* and might reflect that administrative data coding accuracy was impacted by was impacted by the likelihood of greater detail in clinical documentation with severe disease is present, as well as coder experience and expertis.[9] Although it remains unclear why code validity varied throughout the pandemic, it seems reasonable that continuous monitoring of coding validity is needed.[14 15]

The following limitations must be considered when interpreting the research findings. First, while the study presented the sensitivity and PPV, other measures of validity such as specificity and negative predictive value could not be determined because the data could not reliably be used to estimate the true negatives. Thus, evidence on how well the ICD codes perform in excluding COVID-19 cases was not studied in this work. Second, the symptomatic flag in ProvLab is self-reported data voluntarily collected shortly after testing positive, is frequently not available, and is not updated to reflect disease progress, so the results of the corresponding stratified analysis should be interpreted with caution. Third, the PCR test for SARS-CoV-2 may not be a perfect test to constitute a gold standard; however, we chose to use it as it is widely accepted internationally, and is the most practical choice for a large-scale study. Lastly, due to the variability of coding practice and healthcare systems, the generalisability of our findings to other countries or territories or healthcare settings is unknown.

CONCLUSIONS

The validity of ICD-10 code U07.1 and U07.3 demonstrated high sensitivity and PPV in both ED visitors and hospitalised patients. This indicates administrative data in Alberta, Canada, can be used for COVID-19 research and pandemic management purposes.

Twitter Guosong Wu @icd

Contributors HQ conceived this study. YX, GW and AGD'S contributed to the study design. AGD'S and EY retrieved and deidentified the data. AGD'S and GW completed the data analysis. TW, CE and DAS contributed to the data interpretation. GW drafted the manuscript and all authors contributed to the revision. All authors agreed on the final version of submission and account for all aspects of this work. GW is the guarantor and takes responsibility for this work, had access to the data and controlled the decision to publish.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. Due to data sharing policies of the data custodians, the dataset is not able to be made publicly available. It may be able to be shared only to researchers in Alberta with approval from the data custodians.

ORCID iDs
Guosong Wu http://orcid.org/0000-0003-0440-0784
Cathy Eastwood http://orcid.org/0000-0002-4569-8014

REFERENCES

1 Coronavirus (COVID-19) data: World Health organization, 2021. Available: https://www.who.int/data
2 Ozili PK, Arun T. Spillover of COVID-19: impact on the global economy. *SSRN Electronic Journal* 2020;10.
3 Galea S, Merchant RM, Lurie N. The mental health consequences of COVID-19 and physical distancing: the need for prevention and early intervention. *JAMA Intern Med* 2020;180:817–8.
4 Xu B, Gutierrez B, Mekaru S, *et al*. Epidemiological data from the COVID-19 outbreak, real-time case information. *Sci Data* 2020;7:1–6.
5 Weible CM, Nohrstedt D, Cairney P, *et al*. COVID-19 and the policy sciences: initial reactions and perspectives. *Policy Sci* 2020;53:225–41.
6 Hale T, Petherick A, Phillips T. Variation in government responses to COVID-19. *Blavatnik school of government working paper* 2020;31:2020–11.
7 World Health Organization. Emergency use ICD codes for COVID-19 disease outbreak 2020.
8 Information CIfH. ICD-10-CA coding direction for COVID-19. Ottawa; 2021.
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;36:1–14.
10 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.
11 Bodilsen J, Leth S, Nielsen SL. Positive predictive value of ICD-10 diagnosis codes for COVID-19. *Clin Epidemiol* 2021;13:367–72.
12 Quan H, Li B, Couris CM, *et al*. Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge Abstracts using data from 6 countries. *Am J Epidemiol* 2011;173:676–82.
13 Yu AYX, Quan H, McRae AD, *et al*. A cohort study on physician documentation and the accuracy of administrative data coding to improve passive surveillance of transient ischaemic attacks. *BMJ Open* 2017;7:e015234.
14 Santos S, Murphy G, Baxter K, *et al*. Organisational factors affecting the quality of hospital clinical coding. *Health Inf Manag* 2008;37:25–37.
15 Lloyd SS, Rissing JP. Physician and coding errors in patient records. *JAMA* 1985;254:1330–6.