ORIGINAL RESEARCH

Positive Predictive Value of ICD-10 Diagnosis Codes for COVID-19

Jacob Bodilsen [1]
Steffen Leth [2,3]
Stig Lønberg Nielsen [4,5]
Jon Gitz Holler [6]
Thomas Benfield [7]
Lars Haukali Omland [8]

[1]Department of Infectious Diseases, Aalborg University Hospital, Aalborg, Denmark; [2]Department of Infectious Diseases, Aarhus University Hospital Skejby, Aarhus, Denmark; [3]Department of Medicine, Regional Hospital Unit West Jutland, Herning, Denmark; [4]Research Unit for Infectious Diseases, Odense University Hospital, Odense, Denmark; [5]University of Southern Denmark, Odense, Denmark; [6]Department of Pulmonary and Infectious Diseases, Hillerød Hospital, Hillerød, Denmark; [7]Department of Infectious Diseases, Copenhagen University Hospital – Amager and Hvidovre, Hvidovre, Denmark; [8]Department of Infectious Diseases, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark

Purpose: To examine the positive predictive value (PPV) of International Classification version 10 (ICD-10) diagnosis codes for Coronavirus disease 2019 (COVID-19).

Patients and Methods: Medical record review of all patients assigned a diagnosis code of COVID-19 (DB342A or DB972A) at six Danish departments of infectious diseases from February 27 through May 4, 2020. Confirmed COVID-19 diagnosis was defined as either: 1) definite, a positive polymerase chain reaction (PCR) for severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) on a respiratory sample combined with symptoms suggestive of COVID-19: 2) probable, clinical presentation of COVID-19 without detection of SARS-CoV-2 and no alternative diagnoses considered more likely; or 3) possible, clinical presentation of COVID-19 without detection of SARS-CoV-2, and the patient was discharged or deceased before further investigations were carried out. We computed the PPV with 95% confidence intervals (CI) as the number of patients with confirmed (i.e., definite, probable, and possible) COVID-19 divided by the number of patients assigned a diagnosis code for COVID-19.

Results: The study included 710 patients with a median age of 61 years (interquartile range [IQR] 47–74) and 285/710 (40%) were female. COVID-19 was confirmed in 706/710 (99%) with 705/710 (99%) categorized as definite, 1/710 (0.1%) as probable, and 0 patients as possible COVID-19. The diagnosis was disproven in 4/710 (0.6%) patients who were hospitalized due to bacterial pneumonia (n = 2), influenza (n = 1), and urinary tract infection (n = 1). The overall PPV for COVID-19 was 99% (95% CI 99–100) and remained consistently high among all subgroups including sex, age groups, calendar period, and stratified by diagnosis code and department of infectious diseases (range 97–100%).

Conclusion: The overall PPV of diagnosis codes for COVID-19 in Denmark was high and may be suitable for future registry-based prognosis studies of COVID-19.

Keywords: Coronavirus disease 2019; COVID-19, SARS-CoV-2, diagnosis codes, ICD-10, positive predictive value; PPV, validation, epidemiology

Plain-Language Summary

We reviewed the medical records of 710 patients admitted at six departments of infectious diseases in Danish hospitals from February 27 through May 4, 2020 with an ICD-10 diagnosis code of Coronavirus disease (COVID-19) and found an overall positive predictive value (PPV) of 99%. The results were consistent across different diagnosis codes for COVID-19, age groups, sex, calendar period, and departments of infectious diseases. These results are important for ensuring the validity of studies using ICD-10 diagnosis codes to identify COVID-19 patients.

Correspondence: Jacob Bodilsen
Department of Infectious Diseases, Aalborg University Hospital, Mølleparkvej 4, Aalborg, 9000, Denmark
Email Jacob.bodilsen@rn.dk

Introduction

COVID-19 is a potentially life-threatening infection for aging and other vulnerable populations.[1–4] Health care databases allow researchers to effectively explore both common and rare associations in large populations,[5–10] and may be valuable for long-term monitoring of efficiency and safety of treatments and vaccines for COVID-19 outside the settings of randomized controlled trials. In addition, the epidemiological characteristics of COVID-19 are dynamic and rapidly change within and between countries as the pandemic evolves, and large health care registries may assist in quickly testing scientific hypothesis in different settings.[11–15] However, registry-based analyses using diagnosis codes require that the codes are of high quality and, thus far, the validity of ICD-10 codes for COVID-19 remains unclear. This study aimed to examine the PPV of the ICD-10 diagnosis codes for COVID-19 at departments of infectious diseases in Denmark.

Methods
Setting

In Denmark, medical care is tax-supported and free of charge at the point of delivery for all residents. A unique civil registration number assigned at birth or immigration allows for the unique identification of all Danish residents and unambiguous linkage between registries.[7]

Study Design and Study Population

This cross-sectional validation study was conducted by reviewing the medical records of all patients admitted with a first-time ICD-10 diagnosis code for COVID-19 (DB342A and DB972A) at departments of infectious diseases (and affiliated "pandemic departments") at hospitals in Aalborg, Aarhus, Odense, Hillerød, and Copenhagen University Hospitals at Amager/Hvidovre and Rigshospitalet in Denmark from February 27 through May 4, 2020 (Figure 1). Patients were identified by searches at each department and the diagnosis codes correspond to those that are reported to the Danish National Patient Registry.[9]

Record Review and Definition of COVID-19

During medical record review, the admission date of first-time hospitalization for COVID-19 was used as index date. If patients were tested positive for Severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) while hospitalized for other reasons, the date of test for SARS-CoV-2 was considered the index date. Next, a local investigator reviewed the medical records of all hospitalized patients assigned a COVID-19 diagnosis code at each center including doctor's notes, laboratory results, micro-biological analyses, imaging results as described by the

Figure 1 Geographical distribution of the involved departments of infectious diseases in Denmark.

hospital radiologist, and outcomes of the patients. In cases of doubt, categorizations of patients were resolved by discussion (JB and LO). The following definitions for COVID-19 were used:

Definite COVID-19:

1. A positive polymerase chain reaction (PCR) for SARS-CoV-2 on a respiratory sample AND

2. A clinical presentation consistent with COVID-19, e.g., fever, sore throat, headache, nasal congestion, dyspnea, cough, nausea/vomiting, diarrhea, or myalgia.

Probable COVID-19:

1. A clinical presentation consistent with COVID-19, i.e., any combination of fever, sore throat, headache, nasal congestion, dyspnea, cough, nausea/vomiting, diarrhea, and myalgia without detection of SARS-CoV-2 on a respiratory sample AND

2. No other pathogen detected, and no other medical condition was considered more likely.

Possible COVID-19:

1. A clinical presentation consistent with COVID-19, i.e., any combination of fever, sore throat, headache, nasal congestion, dyspnea, cough, nausea/vomiting, diarrhea, and myalgia without detection of SARS-CoV-2 on a respiratory sample AND

2. The patient passed away or was discharged before further examinations had been performed including tests for SARS-CoV-2 and no other medical condition was considered more likely.

A diagnosis of COVID-19 was considered disproven if the patient did not fulfill any of the criteria listed above or had a confirmed alternative diagnosis.

Statistical Analyses

The PPV was calculated as the number of confirmed (i.e., definite, probable, and possible) COVID-19 cases divided by the number of patients assigned a diagnosis code of COVID-19 and using medical record review combined with results of PCR test for SARS-CoV-2 as reference. Using the exact binomial method, the PPV with 95% confidence intervals (CIs) was computed for patients with confirmed COVID-19 and by each category of definite, probable, and possible COVID-19. Additional analyses of the PPV for definite COVID-19 were conducted stratified by diagnosis code (DB342A or DB972A), age groups (0–40, 41–60, 61–80, 81+ years), sex, and calendar

period (February 27 through March 31, and April 1 through May 4). A post hoc analyses considering only definite diagnosis of COVID-19 for assessment of the PPV was also carried out. Categorical variables are presented as n/N (%) and continuous variables as medians with interquartile ranges (IQR).

Study data were collected and managed using REDCap electronic data capture tools hosted at North Denmark Region.[16] Stata/MP® version 16 (StataCorp LLC, Texas) was used for all statistical analyses.

Ethical Considerations

The study was approved by the legal department at the North Denmark Region (record number 2020-045) and the Danish Board of Health (record number 31-1522-84). Patient consent or permission from an ethical committee is not required for this type of study in Denmark. Handling of data complied with relevant data protection and privacy regulations and was conducted in accordance with the Helsinki declaration.

Results

During the study period, a total of 710 patients were assigned a diagnosis code of COVID-19 (Table 1). The median age of patients was 61 years (IQR 47–74) and 285/710 (40%) were females. Consistent with the overall development of the first wave of the pandemic in Denmark, 542/710 (76%) patients were admitted early during the study period from February 27 through March 31, 2020.

A definite diagnosis of COVID-19 was observed in 705/710 (99%) of patients, whereas the diagnosis was probable in 1/710 (0.1%) and possible in 0 patients (Table 2). COVID-19 was disproven in 4/710 (0.6%) patients, of which 2 were diagnosed with bacterial pneumonia (both had unknown pathogens and a clinical response to antibiotic therapy), 1 with influenza, and 1 with a urinary tract infection.

Using both definite and probable diagnoses of COVID-19 as reference, the overall PPV for COVID-19 was 99% (95% CI 99–100) compared with 99% (95% CI 98–100) when using definite cases only (Table 3). The PPV was consistently high and ranged between 97% and 100% among all subgroups including sex, age groups, calendar period, and stratified by diagnosis code (DB342A and DB972A) and department of infectious diseases.

Discussion

This study observed a very high PPV of COVID-19 diagnosis codes for patients hospitalized in Denmark during

Table I Baseline Characteristics of 710 Patients Assigned a First-Time ICD-10 Diagnosis Code for COVID-19 at Departments of Infectious Diseases in Denmark from February 27 Through May 4, 2020

Characteristic	n (%) or Median (IQR)
Diagnosis codes	
DB342A	566 (80)
DB972A	144 (20)
Sex	
Females	285 (40)
Males	425 (60)
Age group, years	61 (47–74)
0–40	116 (16)
41–60	234 (33)
61–80	263 (37)
81-	97 (14)
Calendar period	
February 27 through March 31, 2020	542 (76)
April 1 through May 4, 2020	168 (24)
Department of infectious diseases	
Aalborg	78 (11)
Aarhus	67 (9)
Hillerød	75 (11)
Odense	83 (12)
Hvidovre/Amager	314 (44)
Rigshospitalet	93 (13)

Abbreviation: IQR, interquartile ranges.

Table 2 Distribution of Confirmed COVID-19 Among Patients Assigned ICD-10 Diagnosis Codes for COVID-19 (DB342A and DB972A) at Departments of Infectious Diseases in Denmark

COVID-19	Number of Observations (n/N)	%
Confirmed	706/710	99
Definite	705/710	99
Probable	1/710	0.1
Possible	0	0
Disproved	4/710	0.6

the first wave of the pandemic. The PPV ranged 97–100% in all examined subgroups including sex, age groups, and when stratified by diagnosis code.

The ongoing COVID-19 pandemic has led to an impressive and unprecedented concerted effort of the entire global scientific community to rapidly explore the characteristics and treatment of this infection resulting in more than 100,000 scientific publications on the topic in year 2020.[17] Databases using ICD-10 diagnosis codes may be a valuable tool in

expanding existing knowledge on COVID-19 by analyzing associations in large-scale populations.[7–9,18] However, it remains pivotal that the accuracy of the used diagnosis codes is high. Using SARS-CoV-2 test positivity as reference, a large registry-based study from a US administrative all-payer repository observed a PPV of 92% (95% CI 91–92) for ICD-10 diagnosis codes for COVID-19.[19] In general, the validity of diagnosis codes in Danish health care databases is high,[9] and the results of the current study provide assurance of the usefulness of ICD-10 codes for COVID-19 in Denmark.

Previous registry-based studies have often utilised large microbiological databases of SARS-CoV-2 test-positive and test-negative persons to explore scientific hypothesis related to COVID-19. Strengths of this approach include capture of the majority of individuals with proven infection and identification of a potential control population (i.e., SARS-CoV-2 test negative persons). However, a large proportion of SARS-CoV-2 positive individuals may have asymptomatic infection[20–24] and the current study may thus allow researchers to combine clinical disease characteristics of COVID-19 with documented SARS-CoV-2 infection.

This study has limitations. Selection bias may have been introduced by changes in testing strategies and management of COVID-19 patients. However, all patients admitted to hospital with symptoms of COVID-19 were tested for SARS-CoV-2 throughout the study period in Denmark. In addition, the study included all patients assigned a diagnosis code of COVID-19 at the participating centers representing different geographic regions of Denmark. Ascertainment bias, i.e., researchers examining the medical records were aware of the diagnosis code beforehand and may have favored confirmation of COVID-19 in cases of doubt, was mitigated by accessing all available information in the medical records and adhering to a strict a priori definition of confirmed COVID-19. Moreover, 99% of patients had relevant symptoms and tested positive for SARS-CoV-2 by PCR of a respiratory sample. The study was not able to test the sensitivity, specificity, or completeness of the ICD-10 diagnosis codes for COVID-19, because only patients assigned such diagnosis codes were examined.

The generalizability of the results of this study is improved by the large sample size including all patients with a COVID-19 diagnosis code at participating centers representative of a country with free health care for all residents. Still, coding practices may differ from non-infectious diseases departments and other health care settings, or during later stages of the pandemic.

370
DovePress

Clinical Epidemiology 2021:13

Table 3 Positive Predictive Value (PPV) with 95% Confidence Intervals (CI) of ICD-10 Diagnosis Codes for COVID-19 (DB342A and DB972A) at Departments of Infectious Diseases in Denmark

COVID-19	Number of Observations (n/N)	PPV % (95% CI)
Confirmed overall	706/710	99 (99–100)
Definite cases only	705/710	99 (98–100)
Diagnosis codes		
DB342A	564/566	100 (99–100)
DB972A	142/144	99 (95–100)
Sex		
Females	283/285	99 (97–100)
Males	423/425	100 (98–100)
Age group		
0–40	116/116	100 (97–100)
41–60	233/234	100 (98–100)
61–80	261/263	99 (97–100)
81-	96/97	99 (94–100)
Calendar period		
February 27 through March 31, 2020	540/542	100 (99–100)
April 1 through May 4, 2020	166/168	99 (96–100)
Department of infectious diseases		
Aalborg	76/78	97 (91–100)
Aarhus	67/67	100 (95–100)
Hillerød	75/75	100 (95–100)
Odense	83/83	100 (96–100)
Hvidovre/Amager	313/314	100 (98–100)
Rigshospitalet	92/93	99 (94–100)

In conclusion, the overall PPV of ICD-10 diagnosis codes for COVID-19 was very high and may be suitable for future registry-based prognosis studies of COVID-19.

Abbreviations

CI, Confidence interval; COVID-19, Corona virus disease 2019; ICD-10, International Classification of Diseases versions 10; PPV, Positive predictive value; SARS-CoV-2, Severe acute respiratory syndrome coronavirus 2.

Disclosure

The authors report no conflicts of interest in this work.

References

1. Reilev M, Kristensen KB, Pottegård A, et al. Characteristics and predictors of hospitalization and death in the first 11 122 cases with a positive RT-PCR test for SARS-CoV-2 in Denmark: a nationwide cohort. *Int J Epidemiol*. 2020;49(5):1–14. doi:10.1093/ije/dyaa140
2. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet*. 2020;395(10229):1054–1062. doi:10.1016/S0140-6736(20)30566-3
3. Goyal P, Choi JJ, Pinheiro LC, et al. Clinical characteristics of Covid-19 in New York City. *New Engl J Med*. 2020;382 (24):2372–2374. doi:10.1056/NEJMc2010419
4. Woolf SH, Chapman DA, Lee JH. COVID-19 as the leading cause of death in the United States. *JAMA*. 2021;325(2):123–124. doi:10.1001/jama.2020.24865
5. Sørensen HT, Pedersen L, Jorgensen J, Ehrenstein V. Danish clinical quality databases - an important and untapped resource for clinical research. *Clin Epidemiol*. 2016;8:425–427. doi:10.2147/CLEP. S113265
6. Johannesdottir SA, Horváth-Puhó E, Ehrenstein V, Schmidt M, Pedersen L, Sørensen HT. Existing data sources for clinical epidemiology: the Danish National Database of Reimbursed Prescriptions. *Clin Epidemiol*. 2012;4(1):303–313. doi:10.2147/CLEP.S37587
7. Schmidt M, Pedersen L, Sørensen HT. The Danish Civil Registration System as a tool in epidemiology. *Eur J Epidemiol*. 2014;29 (8):541–549. doi:10.1007/s10654-014-9930-3
8. Schmidt M, Schmidt SAJ, Adelborg K, et al. The Danish health care system and epidemiological research: from health care contacts to database records. *Clin Epidemiol*. 2019;11:563–591. doi:10.2147/CLEP.S179083
9. Schmidt M, Schmidt SAJ, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. *Clin Epidemiol*. 2015;7:449–490. doi:10.2147/CLEP.S91125
10. Frank L. Epidemiology: when an entire country is a cohort. *Science*. 2000;287(5462):2398–2399. doi:10.1126/science.287.5462.2398

11. Pottegård A, Kristensen KB, Reilev M, et al. Existing data sources in clinical epidemiology: the Danish COVID-19 Cohort. *Clin Epidemiol*. 2020;12:875–881. doi:10.2147/CLEP.S257519

12. Lund LC, Kristensen KB, Reilev M, et al. Adverse outcomes and mortality in users of non-steroidal anti-inflammatory drugs who tested positive for SARS-CoV-2: a Danish nationwide cohort study. *Plos Med*. 2020;17(9):e1003308. doi:10.1371/journal.pmed.1003308

13. Pottegård A, Kurz X, Moore N, Christiansen CF, Klungel O. Considerations for pharmacoepidemiological analyses in the SARS-CoV 2 pandemic. *Pharmacoepidemiol Drug Saf*. 2020;29 (8):825–831. doi:10.1002/pds.5029

14. Christiansen CF, Pottegård A, Heide-Jørgensen U, et al. SARS-CoV -2 infection and adverse outcomes in users of ACE inhibitors and angiotensin-receptor blockers: a nationwide case-control and cohort analysis. *Thorax*. 2020. doi:10.1136/thoraxjnl-2020-215768

15. Dalager-Pedersen M, Lund LC, Mariager T, et al. Venous thromboembolism and major bleeding in patients with COVID-19: a nationwide population-based cohort study. *Clin Infect Dis*. 2021; ciab003.

16. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42 (2):377–381. doi:10.1016/j.jbi.2008.08.010

17. Else H. How a torrent of COVID science changed research publishing — in seven charts. *Nature*. 2020;588(7839):553. doi:10.1038/ d41586-020-03564-y

18. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58(4):323–337. doi:10.1016/j.jclinepi.2004.10.012

19. Kadri SS, Gundrum J, Warner S, et al. Uptake and accuracy of the diagnosis code for COVID-19 among US hospitalizations. *JAMA*. 2020;324(24):2553–2554. doi:10.1001/jama.2020.20323

20. Sakurai A, Sasaki T, Kato S, et al. Natural history of asymptomatic SARS-CoV-2 infection. *New Engl J Med*. 2020;383(9):885–886. doi:10.1056/NEJMc2013020

21. Wiersinga WJ, Rhodes A, Cheng AC, Peacock SJ, Prescott HC. Pathophysiology, transmission, diagnosis, and treatment of Coronavirus disease 2019 (COVID-19). *JAMA*. 2020;324 (8):782–793. doi:10.1001/jama.2020.12839

22. Gandhi M, Yokoe DS, Havlir DV. Asymptomatic transmission, the Achilles' Heel of current strategies to control Covid-19. *New Engl J Med*. 2020;382(22):2158–2160. doi:10.1056/NEJMe2009758

23. Rothe C, Schunk M, Sothmann P, et al. Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. *New Engl J Med*. 2020;382(10):970–971. doi:10.1056/NEJMc2001468

24. Oran DP, Topol EJ. Prevalence of asymptomatic SARS-CoV-2 infection: a narrative review. *Ann Intern Med*. 2020;173(5):362–367. doi:10.7326/M20-3012