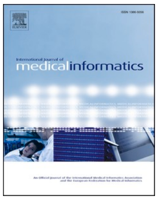
International Journal of Medical Informatics 165 (2022) 104808

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Inaccurate recording of routinely collected data items influences identification of COVID-19 patients

Eva S. Klappe[a,*], Ronald Cornet[a], Dave A. Dongelmans[b], Nicolette F. de Keizer[a]

[a] *Amsterdam UMC, University of Amsterdam, Department of Medical Informatics, Amsterdam Public Health Research Institute, Amsterdam, Netherlands*
[b] *Amsterdam UMC, University of Amsterdam, Department of Intensive Care Medicine, Amsterdam, Netherlands*

ARTICLE INFO

Keywords:
COVID-19
Problem list
Electronic Health Records
Routinely collected data
Data accuracy
Real-time data extraction

ABSTRACT

Background: During the Coronavirus disease 2019 (COVID-19) pandemic it became apparent that it is difficult to extract standardized Electronic Health Record (EHR) data for secondary purposes like public health decision-making. Accurate recording of, for example, standardized diagnosis codes and test results is required to identify all COVID-19 patients. This study aimed to investigate if specific combinations of routinely collected data items for COVID-19 can be used to identify an accurate set of intensive care unit (ICU)-admitted COVID-19 patients.
Methods: The following routinely collected EHR data items to identify COVID-19 patients were evaluated: positive reverse transcription polymerase chain reaction (RT-PCR) test results; problem list codes for COVID-19 registered by healthcare professionals and COVID-19 infection labels. COVID-19 codes registered by clinical coders retrospectively after discharge were also evaluated. A gold standard dataset was created by evaluating two datasets of suspected and confirmed COVID-19-patients admitted to the ICU at a Dutch university hospital between February 2020 and December 2020, of which one set was manually maintained by intensivists and one set was extracted from the EHR by a research data management department. Patients were labeled 'COVID-19′ if their EHR record showed diagnosing COVID-19 during or right before an ICU-admission. Patients were labeled 'non-COVID-19′ if the record indicated no COVID-19, exclusion or only suspicion during or right before an ICU-admission or if COVID-19 was diagnosed and cured during non-ICU episodes of the hospitalization in which an ICU-admission took place. Performance was determined for 37 queries including real-time and retrospective data items. We used the F_1 score, which is the harmonic mean between precision and recall. The gold standard dataset was split into one subset including admissions between February and April and one subset including admissions between May and December to determine accuracy differences.
Results: The total dataset consisted of 402 patients: 196 'COVID-19′ and 206 'non-COVID-19′ patients. F_1 scores of search queries including EHR data items that can be extracted real-time ranged between 0.68 and 0.97 and for search queries including the data item that was retrospectively registered by clinical coders F_1 scores ranged between 0.73 and 0.99. F_1 scores showed no clear pattern in variability between the two time periods.
Conclusions: Our study showed that one cannot rely on individual routinely collected data items such as coded COVID-19 on problem lists to identify all COVID-19 patients. If information is not required real-time, medical coding from clinical coders is most reliable. Researchers should be transparent about their methods used to extract data. To maximize the ability to completely identify all COVID-19 cases alerts for inconsistent data and policies for standardized data capture could enable reliable data reuse.

Abbreviations: CI, Confidence Interval; CORADS, COVID-19 Reporting and Data System; COVID-19, Coronavirus disease 2019; chest CT, chest computerized tomography; DT, Diagnosis Thesaurus; EHR, Electronic Health Record; ICD-10, International Classification of Diseases, Tenth Revision; ICU, intensive care unit; RT-PCR, Reverse-transcription polymerase chain reaction.

* Corresponding author at: Amsterdam UMC, Department of Medical Informatics, Amsterdam, The Netherlands.
E-mail address: e.s.klappe@amsterdamumc.nl (E.S. Klappe).

https://doi.org/10.1016/j.ijmedinf.2022.104808
Received 9 November 2021; Received in revised form 11 April 2022; Accepted 3 June 2022
Available online 10 June 2022

1. Introduction

During pandemics such as the Coronavirus disease 2019 (COVID-19) pandemic, information sharing on patient characteristics, treatment and outcomes is crucial [1–5]. Public health decision-making or forecasting required resources (e.g., ICU beds, ventilators, or protective gear) depends heavily on the number of patients in medical centers [3,5–8]. The hypothesis is that these public health information needs could be fulfilled by reusing Electronic Health Records (EHR) data under the assumption that healthcare professionals keep information on, in this case, COVID-19 patients complete and up-to-date for care purposes, e.g. adjust records when a diagnosis changes from uncertain to confirmed, cured, ruled-out, or when the patient is discharged or deceased. To be able to extract or exchange these data, it is required that these data are stored in a structured and standardized format. Problem lists can help physicians track a patient's status and progress, and organize clinical reasoning and documentation in a structured and standardized way using for instance International Classification of Diseases, Tenth Revision (ICD-10) coding [9–11]. Unfortunately, data in EHRs are highly heterogeneous [12–14] due to variations in unstructured (e.g. free-text) data and incomplete structured data (i.e., current problem lists are not always kept up-to-date) [8,15–18]. Most healthcare professionals believe that free text should always be an option to indicate problems that are hard to code or to indicate uncertainty in diagnoses [19,20]. This suggests that if data are not extracted from appropriate locations in the EHR, or if data are regularly recorded in a free-text field and structured fields are not kept up-to-date, real-time (automatic) extraction will likely produce incomplete or inconsistent information [21–23]. As a result, in the Netherlands, secondary registers for COVID-19 intensive care unit (ICU) admissions were put in place, where data were entered manually by healthcare professionals [24,25]. Manually collected data are considered time-intensive but also error-prone [15,26–28], especially since ICUs were under high pressure [29], which can adversely affect analyses leading to potential erroneous conclusions [27].

While ideally data can be extracted automatically and real-time to support, e.g., public health decision-making, this currently may result in under- or overestimation of the prevalence of patients, which could be a significant hindrance for high-quality research, capacity planning and resource management [3,30–33] as governments take measures based on the numbers reported. To our knowledge, no previous research has systematically investigated the accuracy of routinely collected data for COVID-19 case finding. Hence, the aim of this study is to investigate if specific combinations of routinely collected data items for COVID-19 can be used to identify an accurate set of ICU-admitted COVID-19 patients. We propose recommendations on how to improve data accuracy such that in the future we are better prepared for situations similar to the COVID-19 pandemic that require data collection and processing in real-time thereby also reducing unnecessary administrative workload to record COVID-19 patients twice [6].

2. Material and methods

2.1. Definition of a COVID-19 patient

To better understand what data items are required to accurately identify COVID-19 patients, we need to understand the concept of a 'COVID-19′ patient. The concept 'COVID-19 patient' has been internationally defined as a patient having a positive test result [28] – which is indicated by reverse-transcription polymerase chain reaction (RT-PCR) testing or by chest computed tomography (CT) scans showing COVID-19 Reporting and Data System (CO-RADS) above four [34–36]. However, these tests are not always available, and a patient could have a negative test result but is considered a COVID-19 patient nonetheless due to obvious symptoms and contact with infected cases. The World Health Organization (WHO) has provided specific codes for patients with positive test results irrespective of severity of clinical signs or symptoms (ICD-10 code U07.1) and patients diagnosed clinically or epidemiologically but where laboratory testing is inconclusive or not available (ICD-10 code U07.2) [37,38]. In the Netherlands, the Diagnosis Thesaurus (DT) that underlies problem lists in EHRs includes ICD-10 coded clinical concepts such as U07.1 and U07.2, that are also labeled with synonyms or 'preference' terms. These so-called preference terms for COVID-19 are for instance 'disease caused by sars-cov-2′ (corresponding ICD-10 code: U07.1) or 'disease potentially caused by sars-cov-2′ (corresponding ICD-10 code: U07.2) [39,40]. As described in WHO and Dutch guidelines, U07.2 can therefore be used for (highly) suspected cases of COVID-19 and cases of COVID-19 that are certain, but not confirmed by laboratory testing. In the Netherlands, diagnoses for which the patient was admitted to the hospital are also separately ICD-10-coded by clinical coders (often months) after discharge. Hence, this also applies for COVID-19 patients who were coded retrospectively with U07.1 or U07.2. However, the specific codes for COVID-19 were added and changed over the course of two months which required adjusting codes retrospectively for some patients by healthcare professionals or clinical coders, such as 'other viral pneumonia' that was first advised to use (corresponding ICD-10 code: J12.89) [38,40,41]. Additionally, in our hospital, a specialized infection prevention department provides and updates confirmed and suspected (COVID-19) infection labels to patients and potential need for isolation twice a day. Healthcare professionals can also add infection labels. In conclusion, for this study, we used the four data items to identify a COVID-19 patient: positive RT-PCR test results, COVID-19 coding from healthcare professionals, COVID-19 coding from clinical coders and infection labels.

2.2. Data collection

We performed a retrospective analysis on routinely collected data from two sources including suspected and confirmed COVID-19 patients admitted to the Amsterdam University Medical Center between 1 February 2020 and 31 December 2020:

- *The ICU dataset:* the dataset included clinically confirmed COVID-19 patients and their unique patient identifiers (provided by the hospital) and ICU admission and discharge dates. This list was prospectively and manually maintained outside of the EHR system by intensivists and retrieved by researchers as a single Excel file.
- *The EHR extract dataset:* The data research department of this Dutch university hospital queried the EHR system (Epic) for all confirmed and suspected COVID-19 patients and stored the results in a data warehouse, from which the researchers could retrieve it via a secure server as a single Excel file. The criteria used by the data research department are based on RT-PCR test results, COVID-19 coding from healthcare professionals and infection labels, shown in Appendix A. As a result, the dataset included unique patient identifiers; hospital admission and discharge dates; the previous, current and next wards that indicate departments such as the ICU where patients have been admitted within one hospital admission; (sub)specialties; RT-PCR test results; all ICD-10 diagnoses recorded on the problem list by healthcare professionals; and infection labels.

For each patient in the ICU and EHR extract dataset the data research department enriched the data with the ICD-10 diagnoses retrospectively registered by clinical coders from our hospital.

2.3. Data processing

We created one dataset in which we included all adult patients who were labeled suspected and/or confirmed COVID-19 at any point during their hospital admission from the EHR extract dataset who have also been admitted to the ICU department before 31 December 2020 at some point during their hospital admissions by selecting patient records that

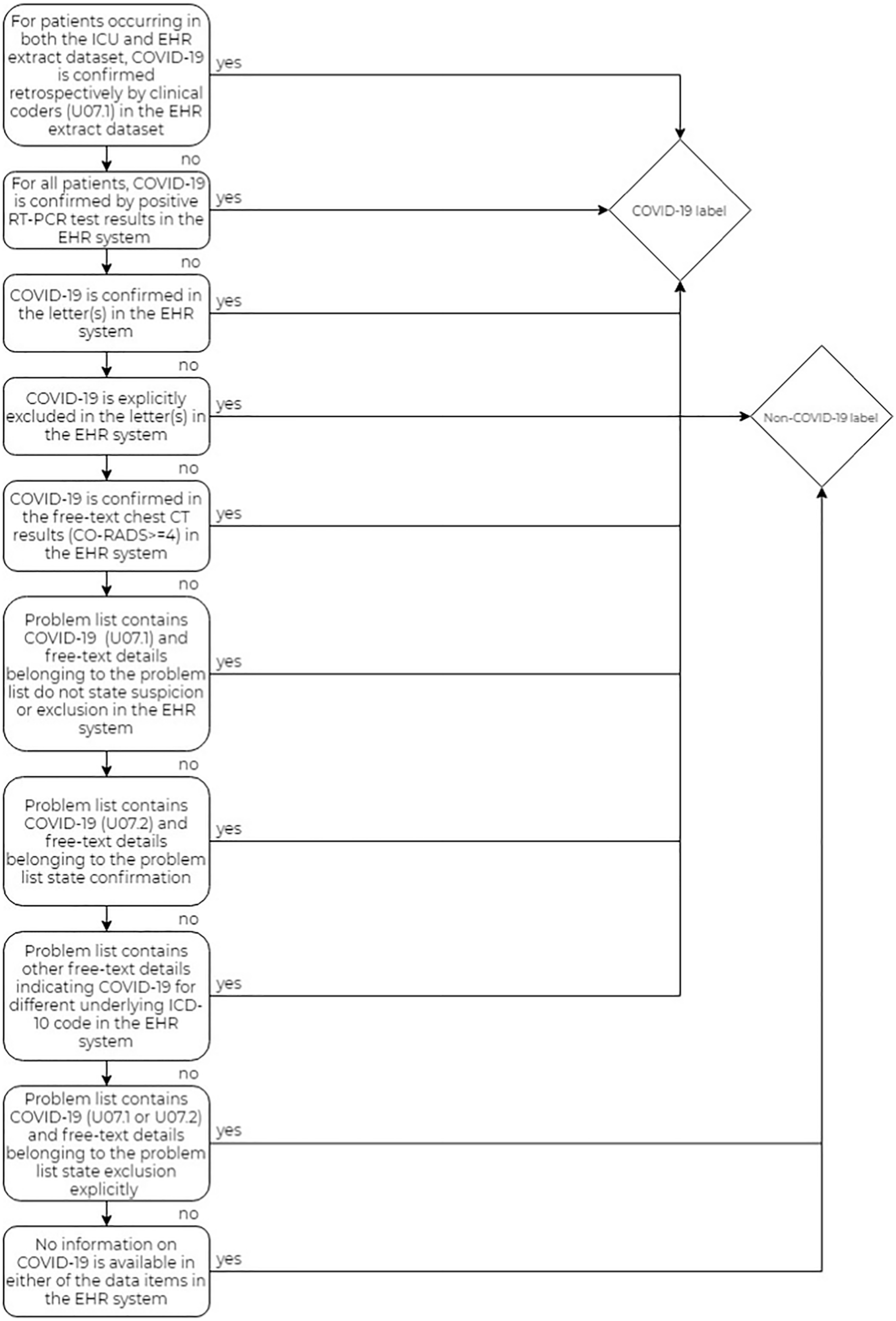

For patients occurring in both the ICU and EHR extract dataset, COVID-19 is confirmed retrospectively by clinical coders (U07.1) in the EHR extract dataset
yes
no
For all patients, COVID-19 is confirmed by positive RT-PCR test results in the EHR system
yes
no
COVID-19 label
COVID-19 is confirmed in the letter(s) in the EHR system
yes
no
COVID-19 is explicitly excluded in the letter(s) in the EHR system
yes
no
Non-COVID-19 label
COVID-19 is confirmed in the free-text chest CT results (CO-RADS>=4) in the EHR system
yes
no
Problem list contains COVID-19 (U07.1) and free-text details belonging to the problem list do not state suspicion or exclusion in the EHR system
yes
no
Problem list contains COVID-19 (U07.2) and free-text details belonging to the problem list state confirmation
yes
no
Problem list contains other free-text details indicating COVID-19 for different underlying ICD-10 code in the EHR system
yes
no
Problem list contains COVID-19 (U07.1 or U07.2) and free-text details belonging to the problem list state exclusion explicitly
yes
no
No information on COVID-19 is available in either of the data items in the EHR system
yes

Fig. 1. Flow chart to annotate a patient with a COVID-19 or non-COVID-19 label.

had 'Intensive care volwassenen' (*English: Intensive care for adults)* as location. We also added patients from the ICU dataset that were admitted to the ICU department before 31 December 2020 and removed duplicate patients.

2.3.1. Gold standard annotation by labeling (non–)COVID-19 patients

We annotated each patient in our dataset with a COVID-19 or non-COVID-19 label based on typical EHR data items that could describe the presence or exclusion of a COVID-19 diagnosis (Fig. 1). Patients that were included in both the EHR extract dataset and ICU dataset were labeled 'COVID-19′ if their admission was provided with an ICD-10 code for confirmed COVID-19 (U07.1) that was registered retrospectively by clinical coders. If these codes were not (yet) available, or if patients only occurred in one of the datasets, author ESK manually checked patients in

Table 1
Search queries including routinely collected data items to identify an accurate set of COVID-19 patients. Search queries shown on white background are EHR data items that could be extracted real-time from the EHR. The search query in italic includes the data item that cannot be extracted real-time as it is retrospectively registered.

Search queries
Positive RT-PCR test result
The ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals *
The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **
An infection label for COVID-19 (confirmed)
*The ICD-10 code for COVID-19 (U07.1) by clinical coders ** *

* According to the WHO definition, both U07.1 and U07.2 indicate COVID-19 patients [37].

** According to the WHO definition [38], according to a (Dutch) manual for using the Diagnosis Thesaurus (DT) for healthcare professionals [40], and according to a (Dutch) manual for clinical coders [42], the ICD-10 code U07.1 is used to indicate a patient confirmed by RT-PCR testing and U07.2 can be used to indicate unconfirmed only suspected COVID-19 patients.

the original EHR system on positive RT-PCR test results, discharge and referral letters, free-text chest CT results on CO-RADS>=4, and problem list codes and notes for confirmed, suspected or excluded COVID-19 diagnoses. Patients with a confirmed COVID-19 diagnosis in (one of) the categories during one or multiple ICU-admission(s) were labeled COVID-19 patient. Patients with exclusion, suspicion or no mentioning of COVID-19 in (one of) the categories during an ICU-admission were labeled 'non-COVID-19 patient'. Patients who were diagnosed with COVID-19 during other non-ICU episodes of a hospitalization in which ICU admission took place, and where COVID-19 was not present during their ICU-admission(s) (i.e. recovered before ICU-admission(s) or diagnosed after ICU-admission(s)), were excluded from analysis. In case of uncertainty, an intensivist (DAD) with full access to the EHR made the final decision to annotate a patient as COVID-19 or non-COVID-19 patient. A final gold standard dataset was created where each patient was labeled 'COVID-19′ or 'non-COVID-19′.

2.3.2. Performance of routinely collected data items to identify COVID-19 patients

Some standardized routinely collected data items are theoretically suitable to identify all COVID-19 patients as they do have a value that is necessary and sufficient to discriminate between (non–)COVID-19 patients. Table 1 shows search queries including routinely collected data items that we applied to the gold standard dataset to determine the percentage of (non–)COVID-19 patients per individual item and specific combinations of two and three data items (e.g. % patients retrieved with positive RT-PCR test results *and* confirmed infection label for COVID-19). A total of 37 search queries including the (combinations of) data items were applied to the dataset. As shown, four search queries included data items that can be extracted from the EHR real-time, and one search query, shown in italic, included a data item that is retrospectively registered by clinical coders and cannot be extracted real-time. It is important to mention that we have only included two search strings with regard to COVID-19 specific ICD-10 coding: "U071 and/or U07.2" (the WHO-definition) and "U07.1" (the Dutch definition). That is, because in the Netherlands the ICD-10 code U07.2 is also used for suspected cases, which makes it difficult to determine whether a patient with only U07.2 is an actual COVID-19 patient or not. Confusion matrices were used to determine the performance of each search query. An example of a confusion matrix is shown in Appendix E. Note that positive RT-PCR test results were used to annotate a patient as a 'COVID-19' patient (Fig. 1), thus automatically leading to zero false-positives in the confusion matrices. Performance was defined in terms of recall, specificity and precision. Recall is a measure of how many of the COVID-19 patients were correctly identified with the data item indicating COVID-19, over all COVID-19 cases in our dataset. Specificity is defined as the proportion of patients that were correctly identified not to have the data item indicating COVID-19 (i.e., true negatives). Precision is a measure of how many patients were correctly identified with COVID-19 (i.e., true positives). We also reported the F_1 score, which is the harmonic mean between precision and recall. An F_1 score lies between zero and one where one indicates perfect precision and recall. RStudio statistical software (v 1.2.1335) for Windows was used for data analysis. Exact binomial 95 % confidence intervals (CI) were calculated for recall, specificity and precision using the 'epi.tests' function from the 'epiR' package. The formulas for the recall, specificity, precision and F_1 score are shown in Appendix E. We split the final gold standard dataset into two equally-sized subsets to determine whether data accuracy differed between earlier months (admission dates between 1 February – 30 April) and later months (admission dates between 1 May – 31 December) of the pandemic.

3. Results

3.1. Gold standard annotation by labeling (non–)COVID-patients

Fig. 2 shows that the gold standard dataset included 402 suspected and confirmed COVID-19 patients who had been at the ICU at some point during an admission between 1 and 2-2020 and 31–12-2020, of which 196 patients were labeled COVID-19, 206 patients were labeled non-COVID-19. As shown, sixteen patients were actual COVID-19 patients, but they were excluded because they had not been at the ICU while being diagnosed with COVID-19, but instead went through COVID-19 during other non-ICU episodes of the same hospital admission in which an ICU admission took place.

3.2. Performance of routinely collected data items to identify COVID-19 patients

Table 2 in Appendix B shows the recall, specificity, precision and F_1 scores for the complete set and the two subsets. The number of patients that were retrieved by applying search queries to the complete gold standard dataset and corresponding F_1 scores are shown in Fig. 3, with the legend showing below. In Appendix C similar figures are shown for both subsets. In the complete gold standard dataset, search queries including data items that can be extracted real-time from the EHR had F_1 scores ranging from 0.68 and 0.97 and returned total numbers of patients ranging from 111 to 327. Search queries including the data item that was retrospectively registered by clinical coders after discharge (ICD-10 code U07.1) had F_1 scores ranging from 0.73 and 0.99 and returned total numbers of patients ranging from 112 to 327. Our results show varying F_1 scores over the 37 search queries and over the two time periods without a clear pattern. Table 3 in Appendix D shows more specific details per data item for the COVID-19 and non-COVID-19 labeled patients showing, e.g., the number of patients coded with U07.2. Confusion matrices to determine the performance per search query are shown in Table 5 in Appendix E.

Legend

Number	Search query
1A	Positive RT-PCR test result
2A	The ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals
3A	The ICD-10 code for COVID-19 (U07.1) by healthcare professionals
4A	An infection label for COVID-19 (confirmed)
5A	Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals
6A	Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1) by healthcare professionals
7A	Positive RT-PCR test result **AND** an infection label for COVID-19
8A	The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **AND** an infection label for COVID-19
9A	

(*continued on next page*)

(*continued*)

Number	Search query
	The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **AND** an infection label for COVID-19
10A	Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals
11A	Positive RT-PCR test result **OR t**he ICD-10 code for COVID-19 (U07.1) by healthcare professionals
12A	Positive RT-PCR test result **OR** an infection label for COVID-19
13A	The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **OR** an infection label for COVID-19
14A	The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **OR** an infection label for COVID-19
15A	Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals **AND** an infection label for COVID-19
16A	Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1) by healthcare professionals **AND** the infection label for COVID-19
17A	Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals **OR** an infection label for COVID-19
18A	Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1) by healthcare professionals **OR** the infection label for COVID-19
1B	The ICD-10 code for COVID-19 (U07.1) by clinical coders
2B	Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders
3B	ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders
4B	The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders
5B	An infection label for COVID-19 **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders
6B	Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders
7B	The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders
8B	The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders
9B	An infection label for COVID-19 **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders
10B	Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders
11B	Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1) by healthcare professionals **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders
12B	Positive RT-PCR test result **AND** an infection label for COVID-19 **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders
13B	The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **AND** an infection label for COVID-19 **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders
14B	The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **AND** an infection label for COVID-19 **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders
15B	Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders
16B	Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1) by healthcare professionals **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders
17B	Positive RT-PCR test result **OR** an infection label for COVID-19 **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders
18B	The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **OR** an infection label for COVID-19 **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders
19B	The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **OR** an infection label for COVID-19 **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders

A: Search queries including EHR data items that could be extracted real-time from the EHR.

B: Search queries including the data item that cannot be extracted real-time as it is retrospectively registered.

4. Discussion

4.1. Principal findings

In this study, we investigated if we could use specific combinations of routinely collected data items to identify an accurate set of ICU-admitted COVID-19 patients. Our results showed that if information is not required to be available real-time, e.g. for retrospective research questions, extracting patients with queries including U07.1-codes registered by clinical coders returns a more accurate set than queries including only real-time data items. Earlier studies also showed high reliability of codes by clinical coders [43,44]. However, real-time data is required for monitoring and forecasting the (national) need for ICU beds, ventilators or protective gear [6]. One of the main findings in this study is that depending on the search query to identify COVID-19 patients (in real-time), patients would be missed or wrongly included which might have negative consequences for, e.g., bed capacity planning and research. While one might use a search query that coincidentally returns the correct number of COVID-19 patients, which may hence result in correct bed capacity planning, the combination of false-positives and true-positives may still impact research findings due to including wrong patient characteristics.

The outcomes of this study also showed that including infection labels in a search query resulted mostly in higher performance, but this can be explained by the fact that infection labels are maintained daily by the infection department team in our hospital. Including ICD-10 coding from problem lists resulted overall in a relatively low performance. We hypothesize that healthcare providers used U07.2 to indicate both confirmed and suspected COVID-19 patients, which explains why performance is lower when including U07.2 in search queries. This can be explained by the fact that in the Netherlands synonym or preference terms from the DT that were linked to U07.2 are described by 'suspected' and 'probable', but according to the WHO U07.2 can also be used for confirmed COVID-19 patients, albeit not proven in laboratory tests. Our study also showed variability in the accuracy of U07.1 coding and U07.1 and/or U07.2 coding by healthcare providers over time, without a clear pattern. This could be partially explained by the fact that concepts for COVID-19 such as U07.2 were added over the course of two months and local implementation rules changed. This required healthcare providers to manually adjust codes for some patients [38,40]. Additionally, the variability can be explained by the fact that COVID-19 cases might have been overestimated at the beginning of the pandemic, shown by the higher number of false positives indicating that more (suspected) cases were registered with U07.1. The use of these codes might therefore not be consistent across different hospitals and countries, which is also supported by findings from a study that investigated the accuracy of COVID-19 specific ICD-10 coding using data from the Mass General Brigham health system (Boston, United States) [45]. This study showed overall lower recall (49.2%) and precision (90%) for the use of U07.1 compared to the recall (82%) and precision (99%) for the use of U07.1 in our study. Furthermore, some financial incentives may promote accurate COVID-19 coding [46]. Researchers showed that these increased problem list accuracy by among others providing salary bonuses which increased the willingness of healthcare providers to change their workflows [47]. In the Netherlands hospitals received additional budget for COVID-19 care based on the number of patients treated, this might have influenced the accuracy of the problem list. Our study also showed that not all patients had positive RT-PCR test results, which could be explained by the fact that RT-PCR tests were not always available, especially at the beginning of the COVID-19 pandemic, which may account for the lower recall of positive RT-PCR test results in the first period of time, or because some COVID-19 patients that were transferred from other hospitals were not tested again in the current hospital and data from the former hospital was not exchanged [18].

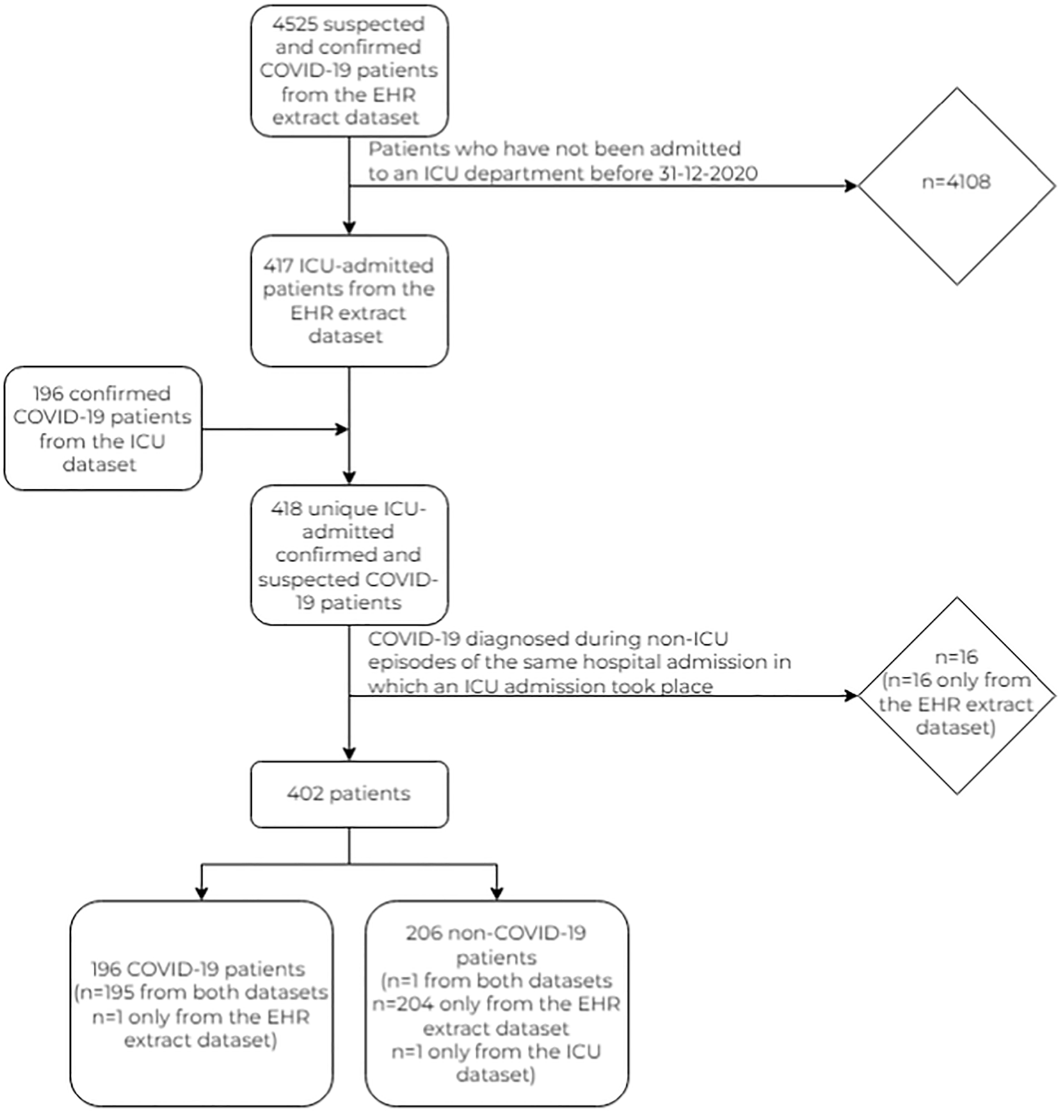

4525 suspected and confirmed COVID-19 patients from the EHR extract dataset
Patients who have not been admitted to an ICU department before 31-12-2020
n=4108
417 ICU-admitted patients from the EHR extract dataset
196 confirmed COVID-19 patients from the ICU dataset
418 unique ICU-admitted confirmed and suspected COVID-19 patients
COVID-19 diagnosed during non-ICU episodes of the same hospital admission in which an ICU admission took place
n=16 (n=16 only from the EHR extract dataset)
402 patients
196 COVID-19 patients (n=195 from both datasets n=1 only from the EHR extract dataset)
206 non-COVID-19 patients (n=1 from both datasets n=204 only from the EHR extract dataset n=1 only from the ICU dataset)

Fig. 2. Dataset inclusion and exclusion and final gold standard dataset (n = 402) with 196 COVID-19 labeled patients and 206 non-COVID-19 labeled patients.

4.2. Relation to other literature

Former studies often used RT-PCR test results to include COVID-19 patients for research [48], sometimes even by including only patients with two positive RT-PCR test results for SARS-CoV-2 [45,49]. Research also shows that chest CT results are considered highly accurate for diagnosing COVID-19 [35], because of good sensitivity [50]. During analysis of patients in the original EHR, chest CT results were included as free text, which made the analysis time-intensive and the results are not interpretable by machines. Considering that one might need all COVID-19 patients for surveillance or bed capacity planning, patients that did not have positive RT-PCR test results but did have positive chest CT results might be missed due to variations in details and the free-text format.

It should be noted that for our COVID-19 use case, identifying patients on testing is possible because disease-specific tests exist. For other diseases, these tests or other markers might be lacking, which makes researchers, governments and other parties more dependent on (standardized) diagnoses on problem lists. Recent studies show that researchers strongly rely on (other) coding systems (ICD-10 and SNOMED CT) to select cohorts for research, for instance whether patients diagnosed with substance use disorders were at increased risk for COVID-19 [51]. Another retrospective study included 513,284 confirmed COVID-19 cases based on “a cohort of all patients who had a confirmed diagnosis of COVID-19 (ICD-10 code U07.1)” [52]. However, this requires that problem list codes should be maintained when new evidence becomes available that proves the existence or absence of the disease. Our current study showed that when using ICD-10 coding from problem lists we would have both wrongly included and missed COVID-19 patients which indicates that problem lists are not kept up-to-date, e.g., old problems are not removed or resolved. This is also in line with previous research [9,19,20,47,53–62]. Research further shows that problem list use varies between specialties [47,63], diseases [64] and between providers [65]. Providers are more likely to update problem lists for first-time patients than for patients they have seen before [65]. We further hypothesize that the accuracy of ICD-10 coding may vary between patients who have died or survived, as the reliability of ICD-10 coded cause of death mentioned on death certificates is variable [66–68]. Although this study does not take into account the impact of specific demographics on coding accuracy, we believe that this should be further investigated for COVID-19 and other diseases.

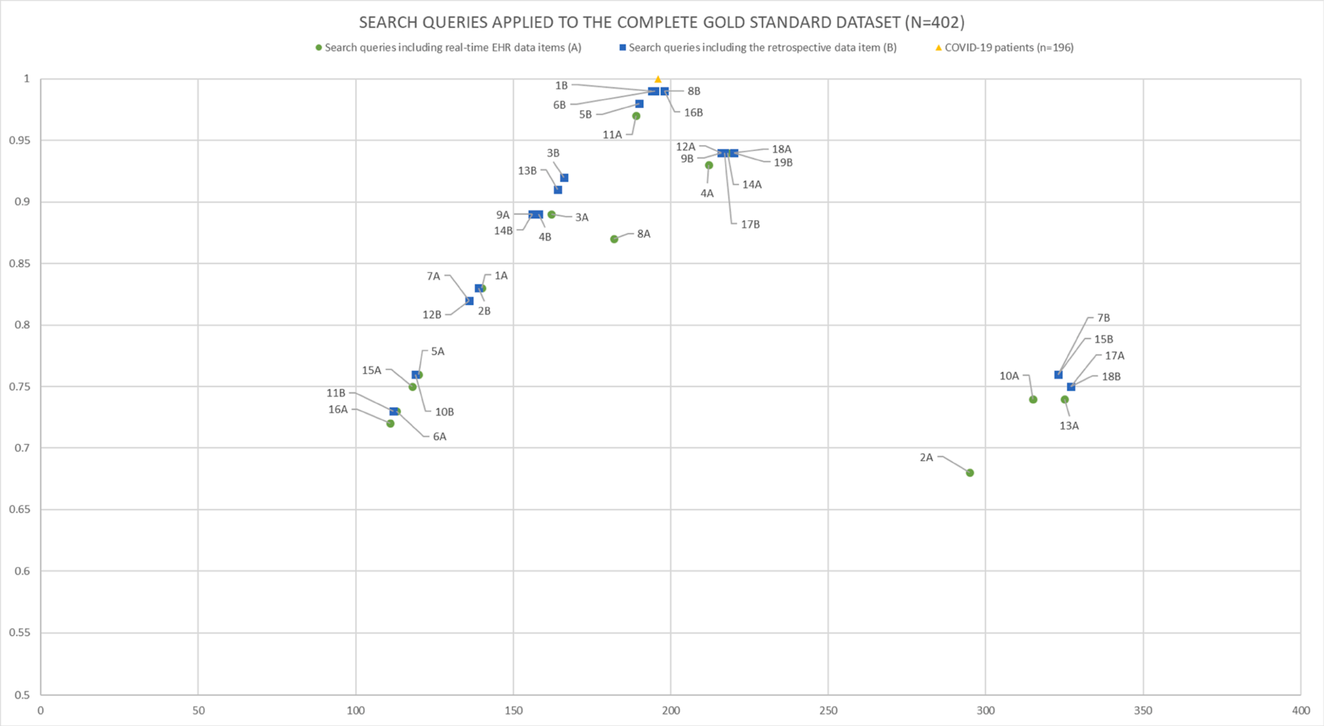


Fig. 3. Search queries applied to the gold standard dataset (n = 402). The numbers indicate the search queries, shown in the legend.

4.3. *Recommendations to maximize the ability to identify an accurate set of COVID-19 patients*

Firstly, considering that different discrepancies might occur when using different search queries, we recommend that researchers should be transparent about their methods of data extraction, which is also supported by recent literature [69]. This also implies that when identifying data items, a clarification of the scope is needed [70], i.e. the specific use case for which the data items are required. For instance, for bed capacity planning a complete cohort of patients is required, but researchers might want to differentiate between patients who have been admitted with COVID-19 (a different primary diagnosis), or due to COVID-19 (COVID-19 being the primary diagnosis). Secondly, it is important to make users aware of benefits (and potential harm to patient care if incorrect) of structured and standardized data capture and encourage better documentation [20,71]. Thirdly, one should be careful using certain (combinations of) data items, particularly when including coded problem list data. Still, evidence suggests that patients with complete problems lists may receive higher quality care than patients with gaps in their problem list [10]. Hence, we believe that a specific policy on keeping a problem list up-to-date, including when to change a working diagnosis (*suspected* covid-19) to the primary diagnosis (*confirmed* COVID-19) and when to close or remove a problem, is essential to reliably reuse problem list data, which is also supported by other studies [9,58,59,62,72]. Fourthly, in a problem-oriented medical record, ordering of RT-PCR tests could require ICD-10 code U07.2 on the problem list. Afterwards, alerts could be implemented in the EHR system to make users aware of this working diagnosis, e.g., a trigger alert for when U07.2 has been on the problem list for more than 24 h. Fifthly, validation rules implemented in the EHR system can be used to identify and solve inconsistencies during care and registration processes. When a patient receives a positive RT-PCR test result, the system could propose the healthcare provider to automatically put ICD-10 code U07.1 on the problem list or update U07.2 to the clinically confirmed ICD-19 code for COVID-19 (U07.1).

4.4. *Strengths and limitations*

Although many COVID-19 studies have been performed based on EHR data, to the best of our knowledge, this was the first study to unravel different routinely collected data items to identify COVID-19 patients. A limitation that should be mentioned is that data were obtained from a single site in the Netherlands but we believe that registration patterns observed in this system resemble those in other hospitals in the Netherlands as well as other western countries with similar system. Hence we believe that hospitals in other countries could learn and benefit from the results as well. Additionally, our study only focused on the accuracy of routinely collected data items for ICU-admitted patients, which could differ from the accuracy of routinely collected data items for patients admitted to the general wards, but not to the ICU [47,63]. Furthermore, in theory we could have missed COVID-19 patients in our gold standard, but we consider this highly unlikely because of the specific attention to COVID-19 in the ICU and research data management department. A potential bias that hampers generalizability of our findings for case finding of other types of patients based on routinely collected EHR data, is that for COVID-19 patient records and specifically problem lists might be kept more accurate than for other diseases, because of higher perceived importance of correctly registering COVID-19 cases. Nonetheless, even for COVID-19, we showed that it is difficult to extract a complete cohort of patients, which is an important finding for future research using the EHR system for data extraction.

5. Conclusions

Our study showed that identifying COVID-19 patients using routinely collected data items can lead to missing or falsely including patients and thus leading to an inaccurate set and incorrect numbers of COVID-19 patients. Researchers should therefore be transparent about their data extraction methods and related limitations. If the reuse purpose of data does not require real-time data, one should consider to include clinical coding by clinical coders after discharge to maximize the ability to completely identify COVID-19 patients. Recommendations to further optimize EHR data quality are among others: the implementation of a problem-oriented structure in the EHR, policy on problem list use, and alerts for inconsistent data. Effectiveness of these recommendations should be evaluated in future research.

Author's contributions

ESK did the analysis of the data and wrote the drafts of the article. RC and NFdK supervised the process and commented on the drafts of the article as presented by ESK. DAD assisted in the final decision on the inclusion or exclusion of some patients for the gold standard.

Funding and declaration of interest

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. This study was funded by Amsterdam UMC 2019-AMC-JK-7. Amsterdam UMC did not have any role in the study design, collection, analysis, interpretation of the data, writing the report and the decision to submit the report for publication.

Statement on conflicts of interest

The authors declare that they have no competing interests.

Statement on author agreement

All authors have read and accepted the final manuscript.

Ethics in publishing

This is a quality improvement project carried out to improve data quality for operational reporting. The project was approved by the data protection officer in the hospital. The Medical Ethics Committee of the Amsterdam UMC judged that this study was not subject to the Dutch Medical Research Involving Human Subjects Act (W20_344 # 20.382) thus the need for ethical approval and patient consent was waived.

Summary table

What was already known on the topic.

- Data in EHRs is highly heterogeneous, which makes it difficult to extract data real-time to guide public health decision-making which was required for the COVID-19 pandemic for e.g. surveillance, bed capacity planning or research.

What this study added to our knowledge.

- The study highlighted that at this point we cannot rely on potential sufficient EHR data items for complete case finding.
- Researchers should be transparent about the methods they used to extract data, and consider using data encoded by clinical coders for more complete case finding.
- Implementation of a problem-oriented structure in the EHR, policies regarding standardized data capture, and alerts for inconsistent data need to be considered to improve data quality in the EHR and to maximize the ability to identify a complete set of COVID-19 patients.

Acknowledgements

The authors would like to thank R. A. de Jong from the Research Data Management department for assistance in creating the EHR extract dataset.

Appendix A

Steps for the query to include COVID-19 patients as suspected or confirmed COVID-19 patients from the EHR into the EHR extract.

1. Select all patient identifiers of patients admitted to the hospital with an infection label for COVID-19 or suspected COVID-19
   a. Suspected COVID-19: patients with suspected COVID-19 or with 'suspected' in the infection label details
   b. Confirmed COVID-19: all other patients
2. Select all patient identifiers of patients admitted to the hospital with an ICD-10 code on the problem list for COVID-19 or suspected COVID-19
   a. Suspected COVID-19: patients with suspected COVID-19
   b. Confirmed COVID-19: all other patients
3. Select all patients admitted to the hospital with RT-PCR test results for 'SarsCov2′ with result: 'positive' or 'follows'
   a. Suspected COVID-19: patients with RT-PCR test result 'follows'
   b. Confirmed COVID-19: patients with RT-PCR test result 'positive'
4. Add all results from the previous steps into one table. This includes duplicative patients.

Appendix B

See Table 2.

Table 2

Performance of search queries including (combinations of) routinely collected data items to identify an accurate set of COVID-19 patients. The performance is determined using the gold standard dataset including the (non–)COVID-19 labels and two subsets. In white, the search queries including data items that could be extracted real-time from the EHR system are shown. In italic, the search queries including ICD-10 coding retrospectively registered by clinical coders are shown.

	Resulting cases (true and false) (n)			Recall (95% CI)			Specificity (95% CI)			Precision (95% CI)			F_1 score		
	Complete set (n = 402, 196 COVID-19; 206 non-COVID-19)	Feb-Apr (n = 208, 90 COVID-19; 118 non-COVID-19)	May-Dec (n = 194, 88 COVID-19; 106 non-COVID-19)	Complete set	Feb-Apr	May-Dec	Complete set	Feb-Apr	May-Dec	Complete set	Feb-Apr	May-Dec	Complete set	Feb-Apr	May-Dec
Average	**198**	**107**	**91**	**0.85 (0.81–0.89)**	**0.91 (0.84–0.94)**	**0.82 (0.76–0.87)**	**0.85 (0.82–0.88)**	**0.77 (0.73–0.80)**	**0.96 (0.91–0.98)**	**0.90 (0.86–0.92)**	**0.85 (0.79–0.88)**	**0.97 (0.91–0.98)**	**0.85**	**0.85**	**0.88**
Positive RT-PCR test result	140	61	79	0.71 (0.65–0.78)	0.68 (0.57–0.77)	0.75 (0.65–0.82)	1.0 (0.98–1.0)	1.0 (0.97–1.00)	1.0 (0.96–1.0)	1.0 (0.97–1.0)	1.0 (0.94–1.0)	1.0 (0.95–1.0)	0.83	0.81	0.85
The ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals	295	200	95	0.86 (0.80–0.90)	0.99 (0.94–1.00)	0.75 (0.65–0.82)	0.38 (0.32–0.45)	0.06 (0.02–0.12)	0.82 (0.72–0.89)	0.57 (0.51–0.63)	0.44 (0.37–0.52)	0.83 (0.74–0.90)	0.68	0.61	0.79
The ICD-10 code for COVID-19 (U07.1) by healthcare professionals	162	88	74	0.82 (0.75–0.87)	0.97 (0.91–0.99)	0.69 (0.59–0.78)	0.99 (0.97–1.00)	0.99 (0.95–1.0)	0.99 (0.94–1.0)	0.99 (0.96–1.00)	0.99 (0.94–1.0)	0.99 (0.93–1.0)	0.89	0.98	0.81
An infection label for COVID-19 (confirmed) by members of the infection department or by healthcare professionals	212	110	102	0.97 (0.93–0.99)	0.99 (0.94–1.0)	0.95 (0.89–0.98)	0.89 (0.84–0.93)	0.82 (0.74–0.89)	0.99 (0.95–1.0)	0.90 (0.85–0.93)	0.81 (0.72–0.88)	0.99 (0.95–1.0)	0.93	0.89	0.97
The ICD-10 code for COVID-19 (U07.1) by clinical coders	*194*	*89*	*105*	*0.99 (0.96–1.0)*	*0.99 (0.94–1.0)*	*0.99 (0.95–1.0)*	*1.0 (0.98–1.0)*	*1.0 (0.97–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.98–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.97–1.0)*	*0.99*	*0.99*	*1.0*
Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals	120	60	60	0.61 (0.54–0.68)	0.67 (0.56–0.76)	0.57 (0.47–0.66)	1.0 (0.98–1.0)	1.0 (0.97–1.0)	1.0 (0.96–1.0)	1.0 (0.97–1.0)	1.0 (0.94–1.0)	1.0 (0.94–1.0)	0.76	0.80	0.72
Positive RT-PCR test result **AND the** ICD-10 code for COVID-19 (U07.1) by healthcare professionals	113	59	54	0.58 (0.50–0.65)	0.66 (0.55–0.75)	0.51 (0.41–0.61)	1.0 (0.98–1.0)	1.0 (0.97–1.0)	1.0 (0.96–1.0)	1.0 (0.97–1.0)	1.0 (0.94–1.0)	1.0 (0.93–1.0)	0.73	0.79	0.68
Positive RT-PCR test result **AND** an infection label for COVID-19	136	60	76	0.69 (0.62–0.76)	0.67 (0.56–0.76)	0.72 (0.62–0.80)	1.0 (0.98–1.0)	1.0 (0.97–1.)	1.0 (0.96–1.0)	1.0 (0.97–1.0)	1.0 (0.94–1.0)	1.0 (0.95–1.0)	0.82	0.80	0.84

(*continued on next page*)

Table 2 (*continued*)

	Resulting cases (true and false) (n)			Recall (95% CI)			Specificity (95% CI)			Precision (95% CI)			F_1 score		
	Complete set (n = 402, 196 COVID-19; 206 non-COVID-19)	Feb-Apr (n = 208, 90 COVID-19; 118 non-COVID-19)	May-Dec (n = 194, 88 COVID-19; 106 non-COVID-19)	Complete set	Feb-Apr	May-Dec	Complete set	Feb-Apr	May-Dec	Complete set	Feb-Apr	May-Dec	Complete set	Feb-Apr	May-Dec
Positive RT-PCR test result ***AND*** *the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*139*	*60*	*79*	*0.71 (0.64–0.77)*	*0.67 (0.56–0.76)*	*0.75 (0.65–0.82)*	*1.0 (0.98–1.0)*	*1.0 (0.97–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.97–1.0)*	*1.0 (0.94–1.0)*	*1.0 (0.95–1.0)*	*0.83*	*0.80*	*0.85*
The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **AND** an infection label for COVID-19	182	106	76	0.84 (0.78–0.89)	0.98 (0.92–1.0)	0.72 (0.62–0.80)	0.91 (0.87–0.95)	0.85 (0.77–0.91)	1.0 (0.96–1.0)	0.90 (0.85–0.94)	0.83 (0.74–0.90)	1.0 (0.95–1.0)	0.87	0.90	0.84
The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals ***AND*** *the ICD-10 code (U07.1) for COVID-19 by clinical coders*	*166*	*88*	*78*	*0.85 (0.79–0.89)*	*0.98 (0.92–1.0)*	*0.74 (0.64–0.82)*	*1.0 (0.98–1.0)*	*1.0 (0.97–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.98–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.95–1.0)*	*0.92*	*0.99*	*0.85*
The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **AND** an infection label for COVID-19	156	86	70	0.80 (0.73–0.85)	0.96 (0.89–0.99)	0.66 (0.56–0.75)	1.0 (0.98–1.0)	1.0 (0.97–1.0)	1.0 (0.96–1.0)	1.0 (0.98–1.0)	1.0 (0.96–1.0)	1.0 (0.95–1.0)	0.89	0.98	0.80
The ICD-10 code for COVID-19 (U07.1) by healthcare professionals ***AND*** *the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*158*	*86*	*72*	*0.81 (0.74–0.86)*	*0.96 (0.89–0.99)*	*0.68 (0.58–0.77)*	*1.0 (0.98–1.0)*	*1.0 (0.97–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.98–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.95–1.0)*	*0.89*	*0.98*	*0.81*
An infection label for COVID-19 ***AND*** *the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*190*	*89*	*101*	*0.97 (0.93–0.99)*	*0.99 (0.94–1.0)*	*0.95 (0.89–0.98)*	*1.0 (0.98–1.0)*	*1.0 (0.97–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.98–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.96–1.0)*	*0.98*	*0.99*	*0.98*
Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals	315	201	114	0.96 (0.92–0.98)	1.0 (0.96–1.0)	0.92 (0.86–0.97)	0.38 (0.32–0.45)	0.06 (0.02–0.12)	0.82 (0.72–0.89)	0.60 (0.54–0.65)	0.45 (0.38–0.52)	0.86 (0.78–0.92)	0.74	0.62	0.89

(*continued on next page*)

Table 2 (*continued*)

	Resulting cases (true and false) (n)			Recall (95% CI)			Specificity (95% CI)			Precision (95% CI)			F_1 score		
	Complete set (n = 402, 196 COVID-19; 206 non-COVID-19)	Feb-Apr (n = 208, 90 COVID-19; 118 non-COVID-19)	May-Dec (n = 194, 88 COVID-19; 106 non-COVID-19)	Complete set	Feb-Apr	May-Dec	Complete set	Feb-Apr	May-Dec	Complete set	Feb-Apr	May-Dec	Complete set	Feb-Apr	May-Dec
Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1) by healthcare professionals	189	90	99	0.95 (0.91–0.98)	0.99 (0.94–1.0)	0.92 (0.86–0.97)	0.99 (0.97–1.0)	0.99 (0.95–1.0)	0.99 (0.94–1.0)	0.99 (0.96–1.0)	0.99- (0.94–1.0)	0.99 (0.95–1.0)	0.97	0.99	0.96
Positive RT-PCR test result **OR** an infection label for COVID-19	216	111	105	0.99 (0.96–1.0)	1.0 (0.96–1.0)	0.98 (0.93–1.0)	0.89 (0.84–0.93)	0.82 (0.74–0.89)	0.99 (0.94–1.0)	0.90 (0.85–0.94)	0.81 (0.73–0.88)	0.99 (0.95–1.0)	0.94	0.90	0.99
Positive RT-PCR test result ***OR*** *the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*195*	*90*	*105*	*0.99 (0.97–1.0)*	*1.0 (0.96–1.0)*	*0.99 (0.95–1.0)*	*1.0 (0.98–1.0)*	*1.0 (0.97–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.98–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.97–1.0)*	*0.99*	*1.0*	*1.0*
The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **OR** an infection label for COVID-19	325	204	121	0.99 (0.96–1.0)	1.0 (0.96–1.0)	0.98 (0.93–1.0)	0.36 (0.30–0.43)	0.03 (0.01–0.08)	0.81 (0.71–0.88)	0.60 (0.54–0.65)	0.44 (0.37–0.51)	0.86 (0.78–0.92)	0.74	0.61	0.92
The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals ***OR*** *the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*323*	*201*	*122*	*1.0 (0.98–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.97–1.0)*	*0.38 (0.32–0.45)*	*0.06 (0.02–0.12)*	*0.82 (0.72–0.89)*	*0.61 (0.55–0.66)*	*0.45 (0.38–0.52)*	*0.87 (0.80–0.92)*	*0.76*	*0.62*	*0.93*
The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **OR** an infection label for COVID-19	218	112	106	0.99 (0.96–1.0)	1.0 (0.96–1.0)	0.98 (0.93–1.0)	0.88 (0.83–0.92)	0.81 (0.73–0.88)	0.98 (0.92–1.0)	0.89 (0.84–0.93)	0.80 (0.72–0.87)	0.98 (0.93–1.0)	0.94	0.89	0.98
The ICD-10 code for COVID-19 (U07.1) by healthcare professionals ***OR*** *the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*198*	*91*	*107*	*1.0 (0.98–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.97–1.0)*	*0.99 (0.97–1.0)*	*0.99 (0.95–1.0)*	*0.99 (0.94–1.0)*	*0.99 (0.96–1.0)*	*0.99 (0.94–1.0)*	*0.99 (0.95–1.0)*	*0.99*	*0.99*	*1.0*

(*continued on next page*)

Table 2 (*continued*)

	Resulting cases (true and false) (n)			Recall (95% CI)			Specificity (95% CI)			Precision (95% CI)			F_1 score		
	Complete set (n = 402, 196 COVID-19; 206 non-COVID-19)	Feb-Apr (n = 208, 90 COVID-19; 118 non-COVID-19)	May-Dec (n = 194, 88 COVID-19; 106 non-COVID-19)	Complete set	Feb-Apr	May-Dec	Complete set	Feb-Apr	May-Dec	Complete set	Feb-Apr	May-Dec	Complete set	Feb-Apr	May-Dec
Infection label for COVID-19 ***OR*** *the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*216*	*110*	*106*	*0.99 (0.96–1.0)*	*0.99 (0.94–1.0)*	*0.99 (0.95–1.0)*	*0.89 (0.84–0.93)*	*0.82 (0.74–0.89)*	*0.99 (0.94–1.0)*	*0.90 (0.85–0.94)*	*0.81 (0.72–0.88)*	*0.99 (0.95–1.0)*	*0.94*	*0.89*	*0.99*
Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals **AND** an infection label for COVID-19	118	59	59	0.60 (0.53–0.67)	0.66 (0.55–0.75)	0.56 (0.46–0.65)	1.0 (0.98–1.0)	1.0 (0.97–1.0)	1.0 (0.96–1.0)	1.0 (0.97–1.0)	1.0 (0.94–1.0)	1.0 (0.94–1.0)	0.75	0.97	0.72
Positive RT-PCR test result ***AND*** *the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals* ***AND*** *the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*119*	*59*	*60*	*0.61 (0.54–0.68)*	*0.66 (0.55–0.75)*	*0.57 (0.47–0.66)*	*1.0 (0.98–1.0)*	*1.0 (0.97–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.97–1.0)*	*1.0 (0.94–1.0)*	*1.0 (0.94–1.0)*	*0.76*	*0.79*	*0.72*
Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1) by healthcare professionals **AND** the infection label for COVID-19	111	58	53	0.57 (0.49–0.64)	0.64 (0.54–0.74)	0.50 (0.40–0.60)	1.0 (0.98–1.0)	1.0 (0.97–1.0)	1.0 (0.96–1.0)	1.0 (0.97–1.0)	1.0 (0.94–1.0)	1.0 (0.93–1.0)	0.72	0.78	0.67
Positive RT-PCR test result ***AND*** *the ICD-10 code for COVID-19 (U07.1) by healthcare professionals* ***AND*** *the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*112*	*58*	*54*	*0.57 (0.50–0.64)*	*0.64 (0.54–0.74)*	*0.51 (0.41–0.61)*	*1.0 (0.98–1.0)*	*1.0 (0.97–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.97–1.0)*	*1.0 (0.94–1.0)*	*1.0 (0.93–1.0)*	*0.73*	*0.78*	*0.68*

(*continued on next page*)

Table 2 (*continued*)

	Resulting cases (true and false) (n)			Recall (95% CI)			Specificity (95% CI)			Precision (95% CI)			F_1 score		
	Complete set (n = 402, 196 COVID-19; 206 non-COVID-19)	Feb-Apr (n = 208, 90 COVID-19; 118 non-COVID-19)	May-Dec (n = 194, 88 COVID-19; 106 non-COVID-19)	Complete set	Feb-Apr	May-Dec	Complete set	Feb-Apr	May-Dec	Complete set	Feb-Apr	May-Dec	Complete set	Feb-Apr	May-Dec
*Positive RT-PCR test result **AND** an infection label for COVID-19 **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*136*	*60*	*76*	*0.69 (0.62–0.76)*	*0.67 (0.56–0.76)*	*0.72 (0.62–0.80)*	*1.0 (0.98–1.0)*	*1.0 (0.97–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.97–1.0)*	*1.0 (0.94–1.0)*	*1.0 (0.95–1.0)*	*0.82*	*0.80*	*0.84*
*The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **AND** an infection label for COVID-19 **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*164*	*88*	*76*	*0.84 (0.78–0.89)*	*0.98 (0.92–1.0)*	*0.72 (0.62–0.80)*	*1.0 (0.98–1.0)*	*1.0 (0.97–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.98–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.95–1.0)*	*0.91*	*0.99*	*0.84*
*The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **AND** an infection label for COVID-19 **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*156*	*86*	*70*	*0.80 (0.73–0.85)*	*0.96 (0.89–0.99)*	*0.66 (0.56–0.75)*	*1.0 (0.98–1.0)*	*1.0 (0.97–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.98–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.95–1.0)*	*0.89*	*0.98*	*0.80*
Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals **OR** an infection label for COVID-19	327	204	123	1.0 (0.98–1.0)	1.0 (0.96–1.0)	1.0 (0.97–1.0)	0.36 (0.30–0.43)	0.03 (0.01–0.08)	0.81 (0.71–0.88)	0.60 (0.54–0.65)	0.44 (0.37–0.51)	0.86 (0.79–0.92)	0.75	0.66	0.93
*Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*323*	*201*	*122*	*1.0 (0.98–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.97–1.0)*	*0.38 (0.32–0.45)*	*0.06 (0.02–0.12)*	*0.82 (0.72–0.89)*	*0.61 (0.55–0.66)*	*0.45 (0.38–0.52)*	*0.87 (0.80–0.92)*	*0.76*	*0.62*	*0.93*
Positive RT-PCR test result **OR** the ICD-10	220	112	108	1.0 (0.98–1.0)	1.0 (0.96–1.0)	1.0 (0.97–1.0)	0.88 (0.83–0.92)	0.81 (0.73–0.88)	0.98 (0.92–1.0)	0.89 (0.84–0.93)	0.80 (0.72–0.87)	0.98 (0.93–1.0)	0.94	0.89	0.99

(*continued on next page*)

Table 2 (*continued*)

	Resulting cases (true and false) (n)			Recall (95% CI)			Specificity (95% CI)			Precision (95% CI)			F_1 score		
	Complete set (n = 402, 196 COVID-19; 206 non-COVID-19)	Feb-Apr (n = 208, 90 COVID-19; 118 non-COVID-19)	May-Dec (n = 194, 88 COVID-19; 106 non-COVID-19)	Complete set	Feb-Apr	May-Dec	Complete set	Feb-Apr	May-Dec	Complete set	Feb-Apr	May-Dec	Complete set	Feb-Apr	May-Dec
code for COVID-19 (U07.1) by healthcare professionals **OR** the infection label for COVID-19															
*Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1) by healthcare professionals **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*198*	*91*	*107*	*1.0 (0.98–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.97–1.0)*	*0.99 (0.97–1.0)*	*0.99 (0.95–1.0)*	*0.99 (0.94–1.0)*	*0.99 (0.96–1.0)*	*0.99 (0.94–1.0)*	*0.99 (0.95–1.0)*	*0.99*	*0.99*	*1.0*
*Positive RT-PCR test result **OR** an infection label for COVID-19 **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*217*	*111*	*106*	*0.99 (0.97–1.0)*	*1.0 (0.96–1.0)*	*0.99 (0.95–1.0)*	*0.89 (0.84–0.93)*	*0.82 (0.74–0.89)*	*0.99 (0.94–1.00)*	*0.90 (0.85–0.94)*	*0.81 (0.73–0.88)*	*0.99 (0.95–1.0)*	*0.94*	*0.90*	*0.99*
*The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **OR** an infection label for COVID-19 **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*327*	*204*	*123*	*1.0 (0.98–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.97–1.0)*	*0.36 (0.30–0.43)*	*0.03 (0.01–0.08)*	*0.81 (0.71–0.88)*	*0.60 (0.54–0.65)*	*0.44 (0.37–0.51)*	*0.86 (0.79–0.92)*	*0.75*	*0.61*	*0.93*
*The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **OR** an infection label for COVID-19 **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*220*	*112*	*108*	*1.0 (0.98–1.0)*	*1.0 (0.96–1.0)*	*1.0 (0.97–1.0)*	*0.88 (0.83–0.92)*	*0.81 (0.73–0.88)*	*0.98 (0.92–1.0)*	*0.89 (0.84–0.93)*	*0.80 (0.72–0.87)*	*0.98 (0.93–1.0)*	*0.94*	*0.89*	*0.99*

Abbreviation: CI, confidence interval.

Appendix C

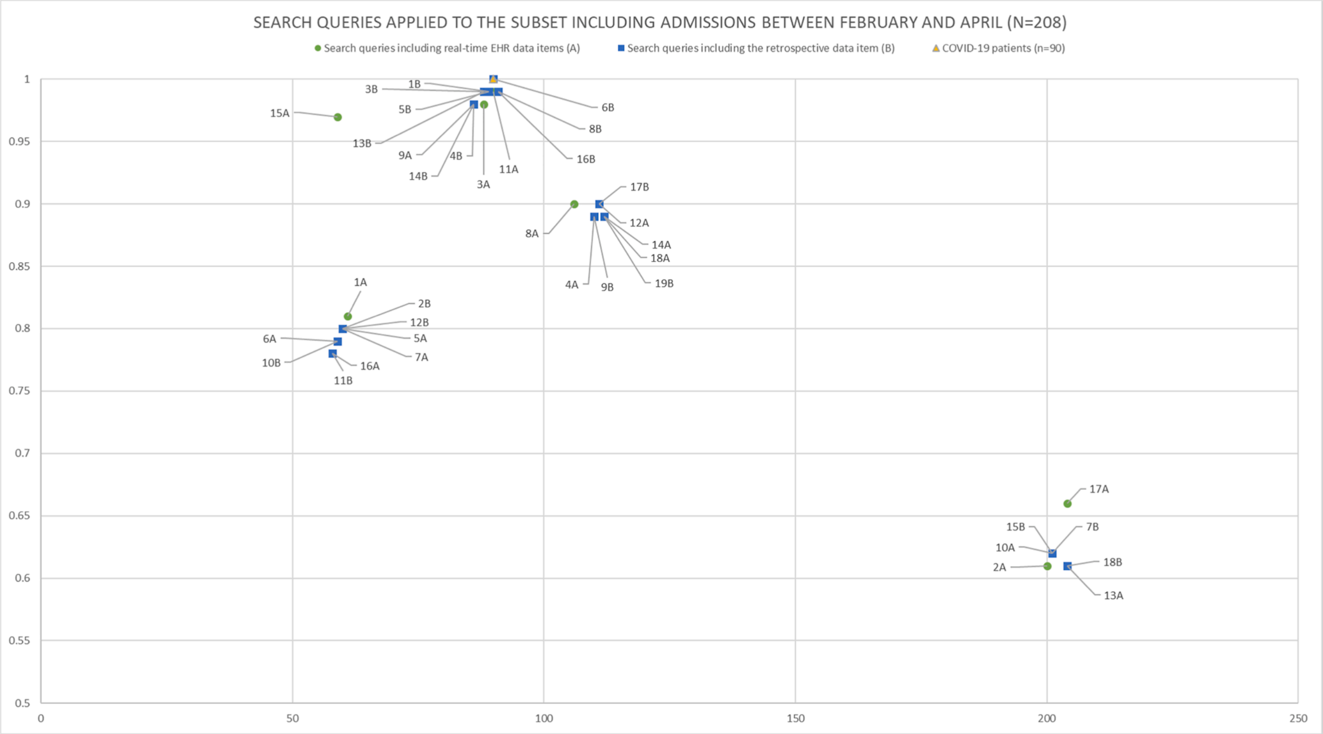


Fig. 4. Search queries applied to the subset including admissions between February and April (n = 208). The numbers indicate the search queries, shown in the legend.

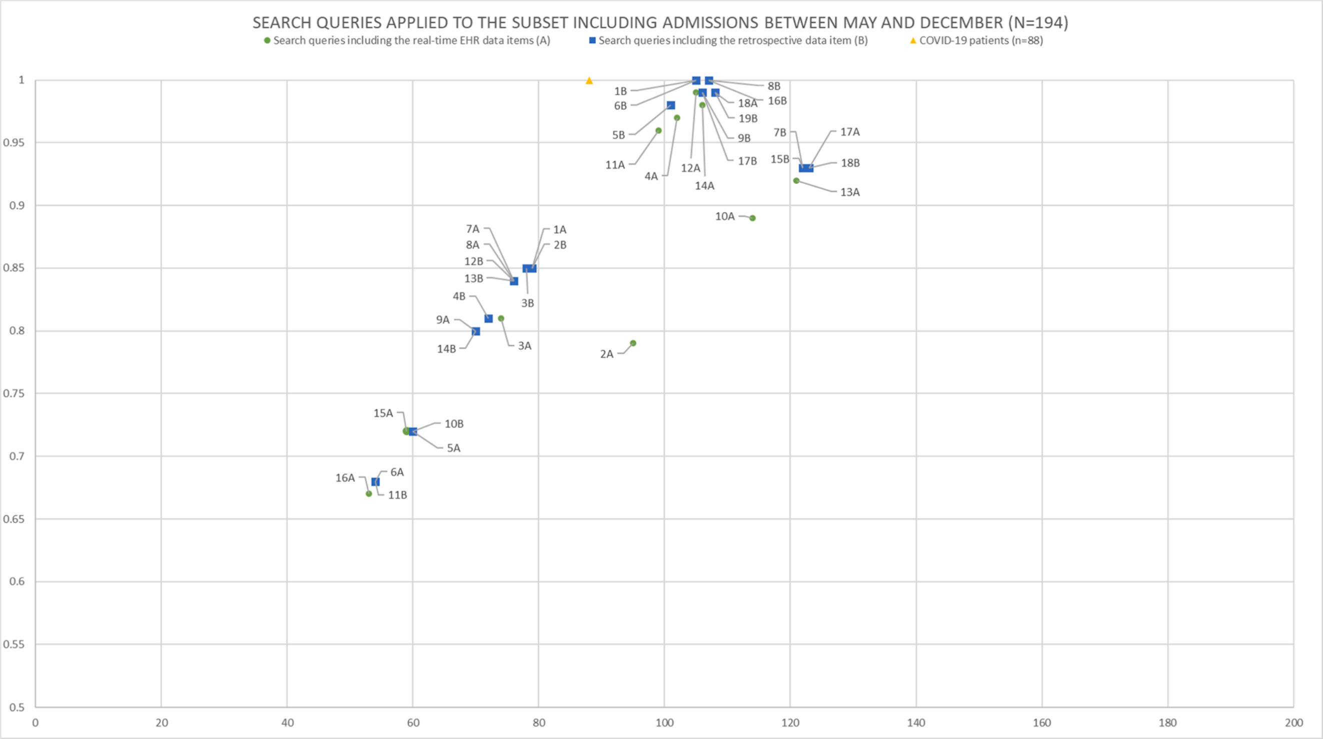


Fig. 5. Search queries applied to the subset including admissions between May and December (n = 194). The numbers indicate the search queries, shown in the legend.

Legend for Fig. 4 and Fig. 5.

Number	Search query
1A	Positive RT-PCR test result
2A	The ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals
3A	The ICD-10 code for COVID-19 (U07.1) by healthcare professionals
4A	An infection label for COVID-19 (confirmed)
5A	Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals
6A	Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1) by healthcare professionals
7A	Positive RT-PCR test result **AND** an infection label for COVID-19
8A	The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **AND** an infection label for COVID-19
9A	The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **AND** an infection label for COVID-19
10A	Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals
11A	Positive RT-PCR test result **OR t**he ICD-10 code for COVID-19 (U07.1) by healthcare professionals
12A	Positive RT-PCR test result **OR** an infection label for COVID-19
13A	The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **OR** an infection label for COVID-19
14A	The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **OR** an infection label for COVID-19
15A	Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals **AND** an infection label for COVID-19
16A	Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1) by healthcare professionals **AND** the infection label for COVID-19
17A	Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals **OR** an infection label for COVID-19
18A	Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1) by healthcare professionals **OR** the infection label for COVID-19
1B	The ICD-10 code for COVID-19 (U07.1) by clinical coders
2B	Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders
3B	ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders
4B	The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders
5B	An infection label for COVID-19 **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders
6B	Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders
7B	The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders
8B	The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders
9B	An infection label for COVID-19 **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders
10B	Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders
11B	Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1) by healthcare professionals **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders
12B	Positive RT-PCR test result **AND** an infection label for COVID-19 **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders
13B	The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **AND** an infection label for COVID-19 **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders
14B	The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **AND** an infection label for COVID-19 **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders
15B	Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders
16B	Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1) by healthcare professionals **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders
17B	Positive RT-PCR test result **OR** an infection label for COVID-19 **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders
18B	The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **OR** an infection label for COVID-19 **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders
19B	The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **OR** an infection label for COVID-19 **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders

A: Search queries including EHR data items that could be extracted real-time from the EHR.

B: Search queries including the data item that cannot be extracted real-time as it is retrospectively registered.

Appendix D

Table 3
Number and percentages of patients retrieved in the complete gold standard dataset (n = 402) based on search queries including routinely collected data items.

Gold standard dataset = 402 patients	COVID-19 patients (n = 196) (n(%))	Non-COVID-19 patients (n = 206) (n(%))
RT-PCR test*		
Confirmed (only ‘positive’)	90 (45.9)	-
Confirmed (Both ‘positive’ and ‘negative’)	50 (25.5)	-
Not-confirmed (only ‘negative’)	18 (9.2)	201 (97.6)
No RT-PCR tests available	23 (11.7)	3 (1.4)
Only other test results (no negative, no positive, not both)	15 (7.7)	2 (1.0)
ICD-10 codes on problem list coded by healthcare professionals**		
U07.1	153 (78.1)	2 (1.0)
Code is ‘closed’	131 (85.6)	2 (100.0)
U07.2	8 (4.1)	125 (60.7)
Code is ‘closed’	7 (87.5)	117 (93.6)
Both U07.1 and U07.2***	7 (3.6)	-
U07.2 was older	6 (85.7)	-
U07.1 was older	1 (14.3)	-
Only other coding (no U07.1, no U07.2)	28 (14.3)	79 (38.2)
Infection labels****		
Confirmed (‘SARS’)	129 (65.8)	18 (8.7)

(*continued on next page*)

Table 3 (*continued*)

Gold standard dataset = 402 patients	COVID-19 patients (n = 196) (n(%))	Non-COVID-19 patients (n = 206) (n(%))
Infection note is suspected	4 (3.1)	17 (94.4)
Suspected ('Suspected SARS')	1 (0.5)	113 (54.6)
Infection note is confirmed	0 (0.0)	2 (1.8)
Both confirmed and suspected***	61 (31.1)	4 (1.9)
Suspected was older	58 (95.1)	2 (50.0)
Confirmed was older	3 (4.9)	2 (50.0)
No infection labels	5 (2.6)	60 (29.0)
Only other infection labels (no SARS, no Suspected SARS)	0 (0.0)	11 (5.3)
ICD-10 codes by clinical coders		
U07.1	194 (99.0)	0 (0.0)
U07.2	2 (1.0)	5 (2.4)
No coding	0 (0.0)	0 (0.0)
Only other coding (no U07.1, no U07.2)	0 (0.0)	201 (97.6)

* Patients who did not have one positive and/or one negative test, but other test results (antibodies, invalid tests, cancelled tests) were considered 'only other test results'. Not-confirmed indicated that patients did not have any positive RT-PCR test result.

** Problem list codes are considered 'active' or 'closed'. Problems are closed when the episode is over, but the problem should still be visible in the problem list (i.e. it will be relevant for medical history). When problems are corrected, they should be removed from the problem list, according to the problem list policy in our hospital.

*** Patients with both confirmed and suspected in either infection labels and problem lists, the dates in 'infection start moment' and 'date of observation' were checked to determine whether confirmed and suspected was older for infection labels and problem lists respectively.

****Infection note is a free-text field indicating more details about the infection status, this displays the number of codes that had contradictory information in the infection note compared to the standardized infection label.

Appendix E

Table 4 shows an example of a confusion matrix. Confusion matrices for the search queries are shown in Table 5.

- Recall: TP / (TP + FN)
- Specificity: TN / (FP + TN)
- Precision: TP / (TP + FP)
- F_1 score: (2 * (precision * recall)) / (precision + recall)

Table 4
Confusion matrix example.

		Gold standard	
		Yes	No
Outcome of the algorithm	Yes	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

Table 5
The confusion matrices and number of patients to determine the performance per search query for the complete gold standard dataset and two subsets. The complete dataset (All) included 402 patients (196 COVID-19; 206 non-COVID-19). The dataset with admissions between February – April 2020 (Feb-Apr) included 208 patients (90 COVID-19; 118 non-COVID-19). The dataset with admissions between May-December (May-Dec) included 194 patients (106 COVID-19; 88 non-COVID-19). In white, the search queries including data items that could be extracted real-time from the EHR system are shown. In italic, the search queries including ICD-10 coding retrospectively registered by clinical coders are shown.

	True Positive (TP) (n)			False Positive (FP) (n)			False Negative (FN) (n)			True Negative (TN) (n)		
	All	Feb-Apr	May-Dec	All	Feb-Apr	May-Dec	All	Feb-Apr	May-Dec	All	Feb-Apr	May-Dec
Positive RT-PCR test result	140	61	79	0	0	0	56	29	27	206	118	88
The ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals	168	89	79	127	111	16	28	1	27	79	7	72
The ICD-10 code for COVID-19 (U07.1) by healthcare professionals	160	87	73	2	1	1	36	3	33	204	117	87
An infection label for COVID-19 (confirmed) by members of the infection department or by healthcare professionals	190	89	101	22	21	1	6	1	5	184	97	87
The ICD-10 code for COVID-19 (U07.1) by clinical coders	*194*	*89*	*105*	*0*	*0*	*0*	*2*	*1*	*1*	*206*	*118*	*88*
Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals	120	60	60	0	0	0	76	30	46	206	118	88
Positive RT-PCR test result **AND t**he ICD-10 code for COVID-19 (U07.1) by healthcare professionals	113	59	54	0	0	0	83	31	52	113	118	88
Positive RT-PCR test result **AND** an infection label for COVID-19	136	60	76	0	0	0	60	30	30	206	118	88
*Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*139*	*60*	*79*	*0*	*0*	*0*	*57*	*30*	*27*	*206*	*118*	*88*
The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **AND** an infection label for COVID-19	164	88	76	18	18	0	32	2	30	188	100	88

(*continued on next page*)

Table 5 (*continued*)

	True Positive (TP) (n)			False Positive (FP) (n)			False Negative (FN) (n)			True Negative (TN) (n)		
	All	Feb-Apr	May-Dec	All	Feb-Apr	May-Dec	All	Feb-Apr	May-Dec	All	Feb-Apr	May-Dec
*The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **AND** the ICD-10 code (U07.1) for COVID-19 by clinical coders*	*166*	*88*	*78*	*0*	*0*	*0*	*30*	*2*	*28*	*206*	*118*	*88*
The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **AND** an infection label for COVID-19	156	86	70	0	0	0	40	4	36	206	118	88
*The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*158*	*86*	*72*	*0*	*0*	*0*	*38*	*4*	*34*	*206*	*118*	*88*
*An infection label for COVID-19 **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*190*	*89*	*101*	*0*	*0*	*0*	*6*	*1*	*5*	*206*	*118*	*88*
Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals	188	90	98	127	111	16	8	0	8	79	7	72
Positive RT-PCR test result **OR t**he ICD-10 code for COVID-19 (U07.1) by healthcare professionals	187	89	98	2	1	1	9	1	8	204	117	87
Positive RT-PCR test result **OR** an infection label for COVID-19	194	90	104	22	21	1	2	0	2	184	97	87
*Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*195*	*90*	*105*	*0*	*0*	*0*	*1*	*0*	*1*	*206*	*118*	*88*
The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **OR** an infection label for COVID-19	194	90	104	131	114	17	2	0	2	75	4	71
*The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*196*	*90*	*106*	*127*	*111*	*16*	*0*	*0*	*0*	*79*	*7*	*72*
The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **OR** an infection label for COVID-19	194	90	104	24	22	2	2	0	2	182	96	86
*The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*196*	*90*	*106*	*2*	*1*	*1*	*0*	*0*	*0*	*204*	*117*	*87*
*Infection label for COVID-19 **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*194*	*89*	*105*	*22*	*21*	*1*	*2*	*1*	*1*	*184*	*97*	*87*
Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals **AND** an infection label for COVID-19	118	59	59	0	0	0	78	31	47	206	118	88
*Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*119*	*59*	*60*	*0*	*0*	*0*	*77*	*31*	*46*	*206*	*118*	*88*
Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1) by healthcare professionals **AND** the infection label for COVID-19	111	58	53	0	0	0	85	32	53	206	118	88
*Positive RT-PCR test result **AND** the ICD-10 code for COVID-19 (U07.1) by healthcare professionals **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*112*	*58*	*54*	*0*	*0*	*0*	*84*	*32*	*52*	*206*	*118*	*88*
*Positive RT-PCR test result **AND** an infection label for COVID-19 **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*136*	*60*	*76*	*0*	*0*	*0*	*60*	*30*	*30*	*206*	*118*	*88*
*The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **AND** an infection label for COVID-19 **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*164*	*88*	*76*	*0*	*0*	*0*	*32*	*2*	*30*	*206*	*118*	*88*
*The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **AND** an infection label for COVID-19 **AND** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*156*	*86*	*70*	*0*	*0*	*0*	*40*	*4*	*36*	*206*	*118*	*88*
Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals **OR** an infection label for COVID-19	196	90	106	131	114	17	0	0	0	75	4	71
*Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1 and/or U07.2) by healthcare professionals **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*196*	*90*	*106*	*127*	*111*	*16*	*0*	*0*	*0*	*79*	*7*	*72*
Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1) by healthcare professionals **OR** the infection label for COVID-19	196	90	106	24	22	2	0	0	0	182	96	86
*Positive RT-PCR test result **OR** the ICD-10 code for COVID-19 (U07.1) by healthcare professionals **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*196*	*90*	*106*	*2*	*1*	*1*	*0*	*0*	*0*	*204*	*117*	*87*
*Positive RT-PCR test result **OR** an infection label for COVID-19 **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*195*	*90*	*105*	*22*	*21*	*1*	*1*	*0*	*1*	*184*	*97*	*87*
*The ICD-10 code (U07.1 and/or U07.2) by healthcare professionals **OR** an infection label for COVID-19 **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*196*	*90*	*106*	*131*	*114*	*17*	*0*	*0*	*0*	*75*	*4*	*71*
*The ICD-10 code for COVID-19 (U07.1) by healthcare professionals **OR** an infection label for COVID-19 **OR** the ICD-10 code for COVID-19 (U07.1) by clinical coders*	*196*	*90*	*106*	*24*	*22*	*2*	*24*	*0*	*0*	*182*	*96*	*86*

Research data for this article

Due to legal regulations were are not allowed to make the datasets publicly available for this study. The authors can be contacted to get more information on the datasets.

References

[1] J.J. Reeves, H.M. Hollandsworth, F.J. Torriani, R. Taplitz, S. Abeles, M. Tai-Seale, et al., Rapid response to COVID-19: health informatics support for outbreak management in an academic health system, J Am Med Inform Assoc. 27 (6) (2020) 853–859.

[2] V. Narayan, P.B. Hoong, S. Chuin, Innovative Use of Health Informatics to Augment Contact Tracing during the COVID19 Pandemic in an Acute Hospital, J Am Med Inform Assoc. (2020).

[3] J.H. Moore, I. Barnett, M.R. Boland, Y. Chen, G. Demiris, G. Gonzalez-Hernandez, D.S. Herman, B.E. Himes, R.A. Hubbard, D. Kim, J.S. Morris, D.L. Mowery, M. D. Ritchie, L.i. Shen, R. Urbanowicz, J.H. Holmes, Ideas for how informaticians can get involved with COVID-19 research, BioData Mining 13 (1) (2020).

[4] A.J. Holmgren, N.C. Apathy, J. Adler-Milstein, Barriers to Hospital Electronic Public Health Reporting and Implications for the COVID-19 Pandemic, J Am Med Inform Assoc. (2020).

[5] M. Wolkewitz, L. Puljak, Methodological challenges of analysing COVID-19 data during the pandemic, BMC Med Res Methodol 20 (1) (2020).

[6] S. Madhavan, L. Bastarache, J.S. Brown, A.J. Butte, D.A. Dorr, P.J. Embi, et al., Use of electronic health records to support a public health response to the COVID-19 pandemic in the United States: a perspective from 15 academic medical centers, J Am Med Inform Assoc. 28 (2) (2021) 393–401.

[7] B.o. Xu, M.U.G. Kraemer, B.o. Xu, B. Gutierrez, S. Mekaru, K. Sewalk, A. Loskill, L. Wang, E. Cohn, S. Hill, A. Zarebski, S. Li, C.-H. Wu, E. Hulland, J. Morgan, S. Scarpino, J. Brownstein, O. Pybus, D. Pigott, M. Kraemer, Open access epidemiological data from the COVID-19 outbreak, Lancet Infect Dis. 20 (5) (2020) 534.

[8] K. Häyrinen, K. Saranto, P. Nykänen, Definition, structure, content, use and impacts of electronic health records: a review of the research literature, Int J Med Inform. 77 (5) (2008) 291–304.

[9] A. Wright, F.L. Maloney, J.C. Feblowitz, Clinician attitudes toward and use of electronic problem lists: a thematic analysis, BMC Med Inform Decis Mak. 11 (36) (2011) 1–10.

[10] D.M. Hartung, J. Hunt, J. Siemienczuk, H. Miller, D.R. Touchette, Clinical implications of an accurate problem list on heart failure treatment, J Gen Intern Med. 20 (2) (2005) 143–147.

[11] D.W. Simborg, B.H. Starfield, S.D. Horn, S.A. Yourtee, Information factors affecting problem follow-up in ambulatory care, Med care. 14 (10) (1976) 848–856.

[12] T. Botsis, G. Hartvigsen, F. Chen, C. Weng, Secondary use of EHR: data quality issues and informatics opportunities, Summit on Translat Bioinforma. 2010 (2010) 1.

[13] J.-F. Diaz-Garelli, R. Strowd, B.J. Wells, T. Ahmed, R. Merrill, U. Topaloglu, Lost in translation: diagnosis records show more inaccuracies after biopsy in oncology care EHRs, AMIA Jt Summits Transl Sci Proc. 2019 (2019) 325.

[14] V.N. O'Reilly-Shah, K.R. Gentry, W. Van Cleve, S.M. Kendale, C.S. Jabaley, D. R. Long, The COVID-19 Pandemic Highlights Shortcomings in US Health Care Informatics Infrastructure: A Call to Action, Anesth Analg 131 (2) (2020) 340–344.

[15] P.-Y. Wu, C.-W. Cheng, C.D. Kaddi, J. Venugopalan, R. Hoffman, M.D. Wang, –Omic and electronic health record big data analytics for precision medicine, IEEE Trans Biomed Eng. 64 (2) (2016) 263–273.

[16] T.N. Ricciardi, M.I. Lieberman, M.G. Kahn, editors. Clinical terminology support for a national ambulatory practice outcomes research network. AMIA Annua Symp Proc; 2005: American Medical Informatics Association.

[17] J.A. Linder, E.O. Kaleba, K.S. Kmetik, Using electronic health records to measure physician performance for acute conditions in primary care: empirical evaluation of the community-acquired pneumonia clinical quality measure set, Med care. (2009) 208–216.

[18] I.S. Kohane, B.J. Aronow, P. Avillach, B.K. Beaulieu-Jones, R. Bellazzi, R. L. Bradford, G.A. Brat, M. Cannataro, J.J. Cimino, N. García-Barrio, N. Gehlenborg, M. Ghassemi, A. Gutiérrez-Sacristán, D.A. Hanauer, J.H. Holmes, C. Hong, J. G. Klann, N.H.W. Loh, Y. Luo, K.D. Mandl, M. Daniar, J.H. Moore, S.N. Murphy, A. Neuraz, K.Y. Ngiam, G.S. Omenn, N. Palmer, L.P. Patel, M. Pedrera-Jiménez, P. Sliz, A.M. South, A.L.M. Tan, D.M. Taylor, B.W. Taylor, C. Torti, A.K. Vallejos, K. B. Wagholikar, G.M. Weber, T. Cai, What every reader should know about studies using electronic health record data but may be afraid to ask, J Med Internet Res. 23 (3) (2021) e22219.

[19] S.M. Simons, F.H. Cillessen, J.A. Hazelzet, Determinants of a successful problem list to support the implementation of the problem-oriented medical record according to recent literature, BMC Med Inform Decis Mak. 16 (102) (2016) 1–9.

[20] E.S. Klappe, N.F. de Keizer, R. Cornet, Factors Influencing Problem List Use in Electronic Health Records—Application of the Unified Theory of Acceptance and Use of Technology, Appl Clin Inform. 11 (03) (2020) 415–426.

[21] P.B. Jensen, L.J. Jensen, S. Brunak, Mining electronic health records: towards better research applications and clinical care, Nat Rev Genet. 13 (6) (2012) 395–405.

[22] K.S. Chan, J.B. Fowles, J.P. Weiner, Electronic health records and the reliability and validity of quality measures: a review of the literature, Med Care Res Rev. 67 (5) (2010) 503–527.

[23] Y. Gao, G.-Y. Cai, W. Fang, H.-Y. Li, S.-Y. Wang, L. Chen, Y. Yu, D. Liu, S. Xu, P.-F. Cui, S.-Q. Zeng, X.-X. Feng, R.-D. Yu, Y.a. Wang, Y. Yuan, X.-F. Jiao, J.-H. Chi, J.-H. Liu, R.-Y. Li, X.u. Zheng, C.-Y. Song, N. Jin, W.-J. Gong, X.-Y. Liu, L. Huang, X. Tian, L. Li, H. Xing, D. Ma, C.-R. Li, F. Ye, Q.-L. Gao, Machine learning based early warning system enables accurate mortality risk prediction for COVID-19, Nat Commun. 11 (1) (2020).

[24] Rijksinstituut voor Volksgezondheid en Milieu. Ontwikkeling COVID-19 in grafieken. 2020. Available at: https://www.rivm.nl/coronavirus-covid-19/grafieken. Accessed 1 October 2020.

[25] A. Milinovich, M.W. Kattan, Extracting and utilizing electronic health data from Epic for research, Ann Transl Med. 6 (3) (2018).

[26] D.F. Sittig, H. Singh, Defining health information technology–related errors: new developments since to err is human, Int Arch Intern Med. 171 (14) (2011) 1281–1284.

[27] M.K.H. Hong, H.H.I. Yao, J.S. Pedersen, J.S. Peters, A.J. Costello, D.G. Murphy, C. M. Hovens, N.M. Corcoran, Error rates in a clinical data repository: lessons from the transition to electronic data transfer—a descriptive study, BMJ open. 3 (5) (2013) e002406.

[28] S.E. Sudat, S.C. Robinson, S. Mudiganti, A. Mani, A.R. Pressman, Mind the clinical-analytic gap: Electronic health records and COVID-19 pandemic response, J Biomed Inform. 116 (2021) 103715.

[29] A.F. Hakimzada, R.A. Green, O.R. Sayan, J. Zhang, V.L. Patel, The nature and occurrence of registration errors in the emergency department, Int J Med Inform. 77 (3) (2008) 169–175.

[30] B.D. Jani, J.P. Pell, D. McGagh, H. Liyanage, D. Kelly, S. de Lusignan, et al., Recording COVID-19 consultations: review of symptoms, risk factors, and proposed SNOMED CT terms, BJGP Open. (2020).

[31] E.S. Grange, E.J. Neil, M. Stoffel, A.P. Singh, E. Tseng, K. Resco-Summers, B. J. Fellner, J.B. Lynch, P.C. Mathias, K. Mauritz-Miller, P.R. Sutton, M.G. Leu, Responding to COVID-19: the UW medicine information technology services experience, Appl Clin Inform. 11 (02) (2020) 265–275.

[32] R. Pryor, C. Atkinson, K. Cooper, M. Doll, E. Godbout, M.P. Stevens, G. Bearman, The electronic medical record and COVID-19: is it up to the challenge? Am J Infect Control 48 (8) (2020) 966–967.

[33] K.M. Skinner, D.R. Miller, E. Lincoln, A. Lee, L.E. Kazis, Concordance between respondent self-reports and medical records for chronic conditions: experience from the Veterans Health Study, J Ambul Care Manage. 28 (2) (2005) 102–110.

[34] S.S. Kadri, J. Gundrum, S. Warner, Z. Cao, A. Babiker, M. Klompas, N. Rosenthal, Uptake and Accuracy of the Diagnosis Code for COVID-19 Among US Hospitalizations, JAMA 324 (24) (2020) 2553.

[35] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, L. Xia, Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases, Radiology 296 (2) (2020) E32–E40.

[36] J. Zhao, Y. Zhang, X. He, P. Xie, Covid-ct-dataset: a ct scan dataset about covid-19. arXiv preprint arXiv:200313865. 2020;490.

[37] World Health Organization. Emergency use ICD codes for COVID-19 disease outbreak. 2020. Available at: https://www.who.int/classifications/classification-of-diseases/emergency-use-icd-codes-for-covid-19-disease-outbreak. Accessed 21 May 2021.

[38] World Health Organization. COVID-19 coding in ICD-10. 2020. Available at: https://www.who.int/classifications/icd/COVID-19-coding-icd10.pdf. Accessed 1 November 2021.

[39] Dutch Hospital Data. T-Rex. 2021. Available at: https://trex.dhd.nl/. Accessed 21 September 2021.

[40] Dutch Hospital Data. Handreiking bij uitlevering nieuwe COVID-19 termen. 2020. Available at: https://www.dhd.nl/over-ons/nieuws/documents/20200528%20Handreiking%20bij%20uitlevering%20nieuwe%20COVID-19%20termen%20in%20DT.pdf. Accessed 15 September 2021.

[41] Centers for Disease Control and Prevention. ICD-10-CM Official Coding Guidelines - Supplement Coding encounters related to COVID-19 Coronavirus Outbreak. 2020. Available at: https://www.cdc.gov/nchs/data/icd/interim-coding-advice-coronavirus-March-2020-final.pdf. Accessed 9 November 2021.

[42] Dutch Hospital Data. Codeadviezen expertgroup ICD-10. 2021. Available at: https://www.dhd.nl/producten-diensten/icd10/Documents/Codeadviezen%20Expertgroep%20ICD-10%20%2001-01-2021.pdf. Accessed 21 September 2021.

[43] M. Daniel Luna, F.G.B. de Quirós, M. Leonardo Garfi, E. Soriano, M. OFlaherty, Reliability of secondary central coding of medical problems in primary care by non medical coders, using the International Classification of Primary Care (ICPC). Medinfo. 2001;10(Pt 2):300.

[44] T. Henderson, J. Shepheard, V. Sundararajan, Quality of diagnosis and procedure coding in ICD-10 administrative data, Med care. (2006) 1011–1019.

[45] A.S. Bhatt, E.E. McElrath, B.L. Claggett, D.L. Bhatt, D.S. Adler, S.D. Solomon, M. Vaduganathan, Accuracy of ICD-10 Diagnostic Codes to Identify COVID-19 Among Hospitalized Patients, J. Gen. Intern. Med. 36 (8) (2021) 2532–2535.

[46] D. Juyal, A. Kumar, S. Pal, S. Thaledi, S. Jauhari, V. Thawani, Medical certification of cause of death during COVID-19 pandemic–a challenging scenario, Journal of Family Medicine and Primary Care. 9 (12) (2020) 5896.

[47] N. Liao, R. Kasick, K. Allen, R. Bode, C. Macias, J. Lee, S. Ramachandran, G. Erdem, Pediatric Inpatient Problem List Review and Accuracy Improvement, Hosp Pediatr. 10 (11) (2020) 941–948.

[48] R.K. Owen, S.P. Conroy, N. Taub, W. Jones, D. Bryden, M. Pareek, et al., Comparing associations between frailty and mortality in hospitalised older adults with or without COVID-19 infection: a retrospective observational study using electronic health records, Age ageing. 50 (2) (2021) 307–316.

[49] B. Yu, X. Li, J. Chen, M. Ouyang, H. Zhang, X. Zhao, L. Tang, Q. Luo, M. Xu, L. Yang, G. Huang, X. Liu, J. Tang, Evaluation of variation in D-dimer levels among COVID-19 and bacterial pneumonia: a retrospective analysis, J thromb thrombolys. 50 (3) (2020) 548–557.

[50] F.K. Lekpa, S.R.S. Njonnou, E. Balti, H.N. Luma, S.P. Choukem, Negative antigen RDT and RT-PCR results do not rule out COVID-19 if clinical suspicion is strong, Lancet Infect Dis 21 (9) (2021) 1209.

[51] Q.Q. Wang, D.C. Kaelber, R. Xu, N.D. Volkow, COVID-19 risk and outcomes in patients with substance use disorders: analyses from electronic health records in the United States, Mol psychiatry. 26 (1) (2021) 30–39.

[52] M. Taquet, M. Husain, J.R. Geddes, S. Luciano, P.J. Harrison, Cerebral venous thrombosis: a retrospective cohort study of 513,284 confirmed COVID-19 cases and a comparison with 489,871 people receiving a COVID-19 mRNA vaccine, Center for Open Science Preprint, 2021.

[53] P.M. Martin, L. Sbaffi, Electronic Health Record and Problem Lists in Leeds, United Kingdom: Variability of general practitioners' views, Health Inform J. (2019) 1–14.

[54] B.-Z. Hose, P. Hoonakker, A. Wooldridge, T. Brazelton III, S. Dean, B. Eithun, J. Fackler, A. Gurses, M. Kelly, J. Kohler, N. McGeorge, J. Ross, D. Rusy, P. Carayon, Physician perceptions of the electronic problem list in pediatric trauma care, Appl Clin Inform. 10 (01) (2019) 113–122.

[55] A. Wright, A.B. McCoy, T.-T. Hickman, D.S. Hilaire, D. Borbolla, W.A. Bowes, W. G. Dixon, D.A. Dorr, M. Krall, S. Malholtra, D.W. Bates, D.F. Sittig, Problem list completeness in electronic health records: a multi-site study and assessment of success factors, Int J Med Inform. 84 (10) (2015) 784–790.

[56] A. Wright, J. Feblowitz, F.L. Maloney, S. Henkin, H. Ramelson, J. Feltman, D. W. Bates, Increasing patient engagement: patients' responses to viewing problem lists online, Appl Clin Inform. 05 (04) (2014) 930–942.

[57] E. Chen, M. Garcia-Webb, An analysis of free-text alcohol use documentation in the electronic health record, Appl Clin Inform. 5 (02) (2014) 402–415.

[58] C. Holmes, The problem list beyond meaningful use: part I: the problems with problem lists, J AHIMA. 82 (2) (2011) 30–33.

[59] A. Wright, J. Pang, J.C. Feblowitz, F.L. Maloney, A.R. Wilcox, H.Z. Ramelson, et al., A method and knowledge base for automated inference of patient problems from structured data in an electronic medical record, J Am Med Inform Assoc. 18 (6) (2011) 859–867.

[60] A. Wright, J. Pang, J.C. Feblowitz, F.L. Maloney, A.R. Wilcox, K.S. McLoughlin, et al., Improving completeness of electronic problem lists through clinical decision support: a randomized, controlled trial, J Am Med Inform Assoc. 19 (4) (2012) 555–561.

[61] D.M. Kaplan, Clear writing, clear thinking and the disappearing art of the problem list, J. Hosp. Med. 2 (4) (2007) 199–202.

[62] C. Holmes, M. Brown, D. St Hilaire, A. Wright, Healthcare provider attitudes towards the problem list in an electronic health record: a mixed-methods qualitative study, BMC Med Inform Decis Mak. 12 (127) (2012).

[63] A. Wright, J. Feblowitz, F.L. Maloney, S. Henkin, D.W. Bates, Use of an electronic problem list by primary care providers and specialists, J. Gen. Intern. Med. 27 (8) (2012) 968–973.

[64] E.-C.-H. Wang, A. Wright, Characterizing outpatient problem list completeness and duplications in the electronic health record, J. Am. Med. Inform. Assoc. 27 (8) (2020) 1190–1197.

[65] J.S. Ancker, L.M. Kern, A. Edwards, S. Nosal, D.M. Stein, D. Hauser, R. Kaushal, How is the electronic health record being used? Use of EHR data to assess physician-level variability in technology use, J. Am. Med. Inform. Assoc. 21 (6) (2014) 1001–1008.

[66] A.V. Gundlapalli, A.M. Lavery, T.K. Boehmer, M.J. Beach, H.T. Walke, P.D. Sutton, R.N. Anderson, Death Certificate-Based ICD-10 Diagnosis Codes for COVID-19 Mortality Surveillance—United States, January–December 2020, Morb. Mortal. Wkly Rep. 70 (14) (2021) 523–527.

[67] J. Ioannidis, Over-and under-estimation of COVID-19 deaths, Eur. J. Epidemiol. 36 (6) (2021) 581–588.

[68] P. Harteloh, K. De Bruin, J. Kardaun, The reliability of cause-of-death coding in The Netherlands, Eur. J. Epidemiol. 25 (8) (2010) 531–538.

[69] A.L. Yin, W.L. Guo, E.T. Sholle, M. Rajan, M.N. Alshak, J.J. Choi, P. Goyal, A. Jabri, H.A. Li, L.C. Pinheiro, G.T. Wehmeyer, M. Weiner, M.M. Safford, T.R. Campion, C. L. Cole, Comparing Automated vs, Int. J. Med. Informatics 157 (2022) 104622.

[70] SNOMED International. COVID-19 Data Coding using SNOMED CT v1.2. 2021 updated 19 August 2021. Available at: https://confluence.ihtsdotools.org/display/DOCCV19/COVID-19+Data+Coding+using+SNOMED+CT. Accessed 4 February 2022.

[71] E.S. Klappe, F.J. van Putten, N.F. de Keizer, R. Cornet, Contextual property detection in Dutch diagnosis descriptions for uncertainty, laterality and temporality, BMC Med Inform Decis Mak. 21 (1) (2021) 1–17.

[72] C. Holmes, The problem list beyond meaningful use: part 2: fixing the problem list, J AHIMA. 82 (3) (2011) 32–35.