DANISH MEDICAL JOURNAL

Original Article

Dan Med J 2020;67(9):A04200237

Estimation of the diagnostic accuracy of real-time reverse transcription quantitative polymerase chain reaction for SARS-CoV-2 using re-analysis of published data

Henrik Frank Lorentzen[1], Sigrun A. J. Schmidt[1, 2], Håkon Sandholdt[3] & Thomas Benfield[3, 4]

1) Department of Dermatology, Aarhus University Hospital, 2) Department of Clinical Epidemiology, Aarhus University Hospital, 3) Department of Infectious Diseases, Amager-Hvidovre Hospital, Hvidovre, 4) Institute of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark

Dan Med J 2020;67(9):A04200237

ABSTRACT

INTRODUCTION: As the coronavirus disease 2019 (COVID-19) epidemic evolves and test strategies change, understanding the concepts of testing (gold standard and test performance measures) becomes essential. The challenge of any novel disease is that the gold standard has yet to be defined.

METHODS: We reanalysed published data on real-time reverse transcription quantitative polymerase chain reaction (RT-qPCR) of severe acute respiratory syndrome coronavirus-2 to illustrate how predictive values vary with disease prevalence, sensitivity (set to values between 30% and 95%) and specificity (set to 99% or 99.98%). We used published data on chest CT and RT-qPCR to examine the potential of latent class analysis to estimate the sensitivity and specificity of RT-qPCR when no single gold standard exists.

RESULTS: For the various sensitivity values, the negative predictive value of a RT-qPCR test remained above 92% until a COVID-19 prevalence of > 10%. The positive predictive value (PPV) was more variable. For a sensitivity of 95% and a specificity of 99%, the PPV was < 10% at a prevalence of 0.1%, increasing to about 90% at a prevalence of 10%. This improved to a PPV of 85% and almost 100%, respectively, when specificity increased to 99.98%. In a restricted latent class analysis, the sensitivity was 97.1% and the specificity was 99.9%, which is similar to figures from the Danish Health Authority. However, derived predictive values depended on model specification.

CONCLUSIONS: A high risk of false-positives should be considered when extending the testing strategy, whereas false-negatives may occur during local outbreaks. This may have consequences for, e.g., containment strategies and research. A confirmatory test (e.g., demonstrating seroconversion or repeated RT-qPCR) may be warranted.

FUNDING: none.
TRIAL REGISTRATION: not relevant.

In Denmark, the first case of severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2)-related disease 2019 (COVID-19) was diagnosed on 27 February 2020. The epidemic was initially countered by a containment strategy that was subsequently transitioned into a mitigation strategy [1]. With this shift came a change in the test strategy from an initial focus on potentially infected travellers from high-risk countries to testing persons with severe symptoms and symptomatic persons from vulnerable groups or persons holding critical societal functions, and then to testing persons with mild symptoms. Most recently, the Danish Ministry of Health began offering testing to all adults, even if symptom free. The early prioritised strategy was supported by a limited test capacity only. The standard method used to diagnose COVID-19 is real-time reverse transcription quantitative polymerase chain reaction (RT-qPCR) of SARS-CoV-2 RNA in respiratory samples [2]. Detection of antibodies to SARS-CoV-2 in plasma or serum may indicate past or present exposure to SARS-CoV-2 infection [2].

One of many challenges of any novel disease is that initially the absence of a gold standard against which to evaluate tests. A negative RT-qPCR for SARS-CoV-2 is therefore often interpreted as “disease free”. However, RT-qPCR may lead to false-negative findings due to insufficient and unrepresentative sampling. As the COVID-19 epidemic evolves and the test strategy changes, it is important to understand the basic concepts of test performance, including the gold standard selected and measures of sensitivity, specificity and predictive values. Unfortunately, misconceptions regarding test performance exist; in particular, sensitivity and positive predictive values (PPVs) are often confused. The distinction is, however, of great importance because predictive values depend on disease prevalence and testing has gradually extended to low-prevalence populations.

In this study, we re-analysed published data on RT-qPCR to diagnose COVID-19 in order to illustrate i) how predictive values of RT-qPCR depend on disease prevalence, sensitivity and specificity and ii) how so-called latent class analysis (LCA) might be used to estimate the sensitivity and specificity of multiple clinical or paraclinical tests, e.g. RT-qPCR, where a gold standard is lacking. We use these results to discuss challenges in diagnosing COVID-19 and the potential clinical and societal implications.

METHODS

Estimation of predictive values for real-time reverse transcription quantitative polymerase chain reaction

The sensitivity (Se) of a test is the conditional probability of a positive test result given presence of the disease (the percentage of diseased individuals identified by the test), whereas specificity (Sp) is the conditional probability of a negative test result given absence of disease. The PPV is defined as the conditional probability of having a disease given a positive test (the percentage of individuals with a positive test who truly have the disease). The negative predictive value (NPV) is the conditional probability of being without the disease given a negative test result (the percentage of individuals with a negative test who do *not* have the disease.) While the sensitivity and specificity are characteristics of the test, the predictive values measure the clinical relevance of a test result. They are calculated as:

$$\text{PPV} = \frac{\text{Se} \times \text{prevalence}}{(\text{Se} \times \text{prevalence}) + (1 - \text{Sp})(1 - \text{prevalence})}$$

$$\text{NPV} = \frac{\text{Sp} \times (1 - \text{prevalence})}{\text{Sp} \times (1 - \text{prevalence}) + (1 - \text{Se}) \times \text{prevalence}}$$

Using these equations, we generated plots for the PPV and NPV of RT-qPCR for SARS-CoV-2 infection at different levels of COVID-19 prevalence. For the main analysis, we used the sensitivity (95%) and specificity (99%) for RT-qPCR, as tabled by the Danish Health Authority on 14 April 2020 [3]. However, because a sensitivity down to 30% has been reported depending on the site of sampling [4], we also repeated the analyses for sensitivity ranging from 30% to 80%. Moreover, we set specificity to 99% – the lower level suggested by the Danish Health Authority. However, this figure may be an underestimate. Cross-reactivity to other endemic respiratory viruses has not been found under reference conditions [5]. Contamination etc. are minimised by strict procedures in clinical practice. We therefore also repeated the analyses using a higher specificity of 99.98%, which was also supported by our LCA analysis (see below).

Concept and application of latent class analysis

We examined the potential of LCA for estimating the sensitivity and specificity of RT-qPCR not otherwise available in the beginning of the epidemic. LCA is a statistical method where latent classes are constructed as a proxy for the true but unknown disease status of the individuals. LCA combines information on multiple observed variables (e.g., different diagnostic test results) to group persons with similar distributions into an unobserved "latent class" (i.e., based on conditional probabilities) [6]. In other words, LCA uncovers hidden groups in a dataset, e.g., groups of different risk-accepting behaviour or disease subgroups. Each subgroup (latent class) is unique, but individuals within a subgroup are similar (homogenous). The latent classes are constructed by numerous iterations for establishing the maximum likelihood of a model given the observed data [7]. Each latent class exhibits local independence (i.e., is homogenous) and is defined by its size (π[latent class i] and by the conditional probabilities of an observable variable (π[manifest class j&;latent class i]). The correct number of classes can be assessed using various methods, including Akaike's information criterion (AIC) and the Bayesian information criterion (BIC), to ensure the most appropriate fit (the model with the lowest AIC or BIC) [7, 8]. A hypothetical example illustrating the concept of LCA is available in Supplementary Methods 1.

We estimated the sensitivity and specificity of RT-qPCR to diagnose COVID-19 by applying LCA to test results for chest CT and RT-qPCR, reported in a cross-sectional study conducted by Ai et al in Wuhan, China from 6 January to 6 February 2020 [9]. The purpose of their study was to examine if chest CT may provide a relevant supplement in diagnosing COVID-19. The study included 1,014 patients suspected of COVID-19 who had both chest CT and RT-qPCR recorded. Serial scans and assays were assessed when available. TaqMan One-Step RT-qPCR kits approved by the China Food and Drug Administration were used. The test results are presented in **Table 1** (reproduced from Figure 1 in the study by Ai et al).

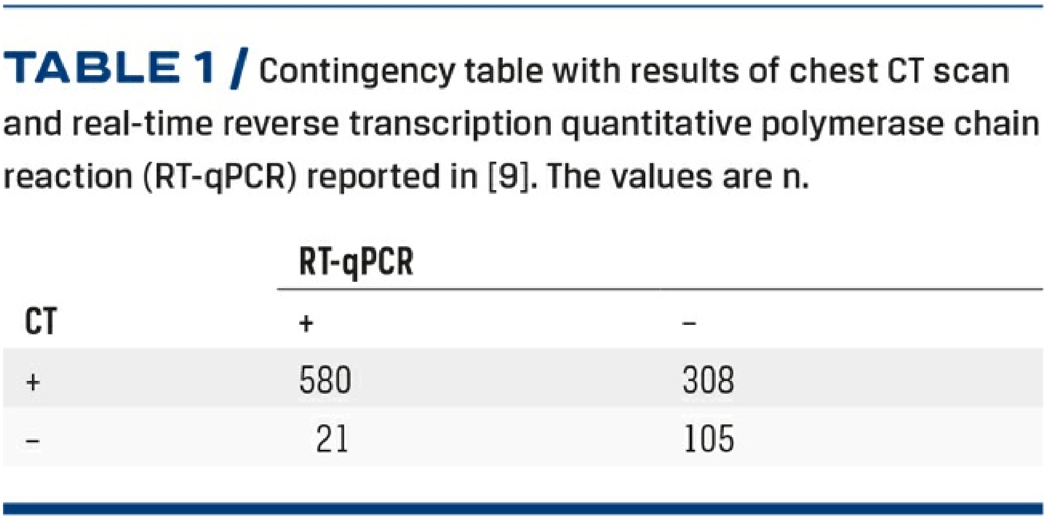


TABLE 1 / Contingency table with results of chest CT scan and real-time reverse transcription quantitative polymerase chain reaction (RT-qPCR) reported in [9]. The values are n.

	RT-qPCR	
CT	+	–
+	580	308
–	21	105

We first used unrestricted LCA to differentiate between two latent classes: I) "COV+" with characteristic (highly probable) COVID-19 patients and II) "COV–" with non-COVID-19 individuals. In the optimal scenario, we would expect the latent class "COV+" to include truly infected persons (concomitant positive RT-qPCR and chest CT), whereas "COV–" would include truly uninfected persons (concomitant negative RT-qPCR and chest CT). The choice of a model with two latent classes was based on a comparison of the AIC, showing potential overfit of a model with a higher number of classes (AIC of ten for three vs four for two latent classes). Recalling the aforementioned definitions of sensitivity and specificity, the conditional probability of a positive RT-qPCR test within the latent class "COV+" (π(RT-qPCR+&;COV+])) would serve as an estimate of the sensitivity for the RT-qPCR test, whereas the conditional probability of a negative RT-qPCR within the latent class of "COV–" (π(RT-qPCR&;COV–])) is an estimate of its specificity.

Unrestricted LCA with only two observable parameters may result in poor model definition, and it has therefore been recommended to add scientifically based constraint(s) on the model to overcome this issue [10]. The unrestricted latent class model can be considered temporary in such situations. We therefore imposed a restriction on the false-positive rate for RT-qPCR, which is the conditional probability of RT-qPCR positivity for the latent class "COV–" (+RT-qPCR&;COV–). We set the start value at 0.01% in the iterative process for the restriction.

We used the sensitivity and specificity for RT-qPCR obtained from the LCA to estimate predictive values by COVID-19 prevalence and compared them with values from the main analysis. Of note, we had originally based all plots on our LCA. However, because of concerns about vulnerable model definitions with only two variables and because new data on the sensitivity and specificity of RT-qPCR emerged during the review process, we based the final plots on the performance estimates from the Danish Health Authority.

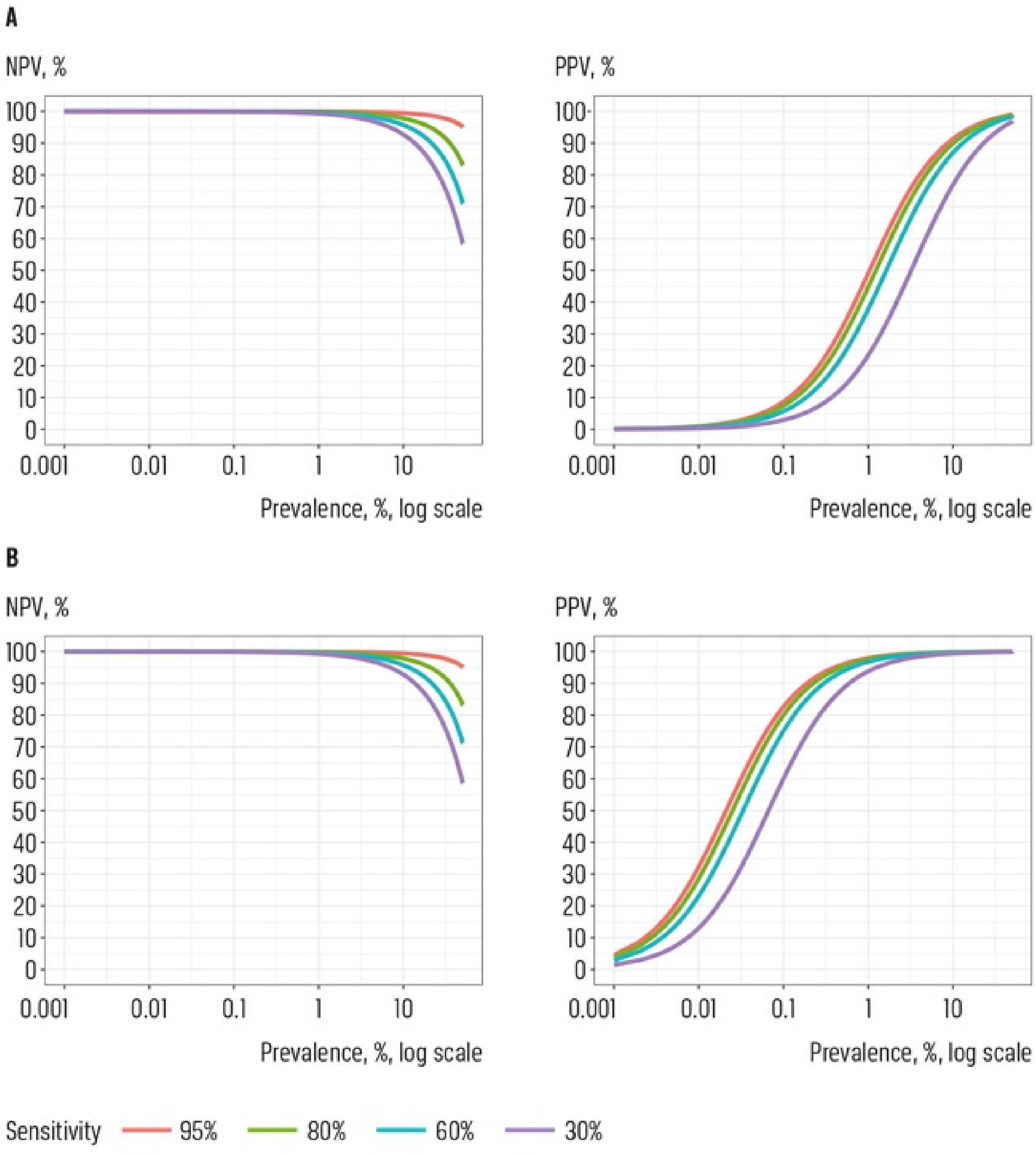
We used the computer programme lEM developed by Vermunt [7, 11] for the LCA (syntax available in Supplementary Methods 2), the MedCalc computer software for calculating predictive values [12] and R statistics version 3.6 for other calculations and to produce graphs [13].

Trial registration: not relevant.

RESULTS

Figure 1A shows the predictive values as a function of prevalence and sensitivity for a specificity of 99%. For the different sensitivities used, the NPV for RT-qPCR remained above 92% until reaching a COVID-19 prevalence above 10%. However, the PPV varied largely for a prevalence between 0.1% (PPV 5-10%) and 10% (PPV 70-90%). Thus, even in the situation with a high sensitivity of 95%, the PPV varies from below 10% at a prevalence of 0.1% to approximately 90% at a prevalence of 10%. For a higher specificity of 99.98% (Figure 1B), the NPVs are largely unchanged, but the PPV improved substantially, varying from 85% at a prevalence of 0.1% to close to 100% at a prevalence of 10%.

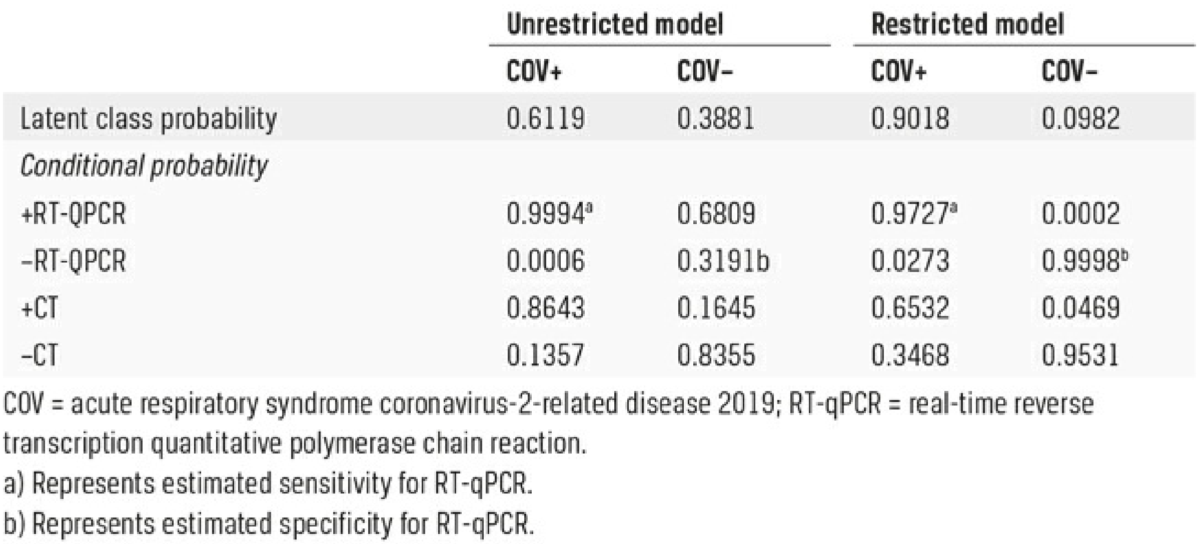
FIGURE 1 / Predictive values of real-time reverse transcription quantitative polymerase chain reaction as a function the prevalence of acute respiratory syndrome coronavirus-2-related disease 2019. **A.** Based on a specificity of 99%, as reported by the Danish Health Authority. **B.** Based on a specificity of 99.98% from the restricted latent class model.



NPV = negative predictive value; PPV = positive predictive value.

For the latent class “COV+” (61% of cases) from the unrestricted LCA, probabilities of a positive RT-qPCR test and chest CT with characteristic COVID-19 findings were 100% and 86%, respectively (**Table 2**). Although the latent class “COV–” (39% of cases) had lower conditional probability of 16% for a positive chest CT, a positive RT-qPCR test was present in 68%, suggesting poor model definition. Conditional probabilities approached expected values in the restricted LCA (Table 2), except for a lower conditional probability for positive chest CT in the “COV+” class (65%). We favoured the restricted model rather than the unrestricted because of its slightly better fit (AIC = 2,029 vs 2,031; BIC 2,048 vs 2,055). The model estimated 99.98% specificity and 97.1% sensitivity for RT-qPCR (Table 2). Overall, these estimates are similar to those reported by the Danish Health Authority (**Table 3**); however, the PPV increased substantially with decreasing start value for the false-positive rate for RT-qPCR in the iterative process.

TABLE 2 / Unrestricted and restricted models from latent class analysis.



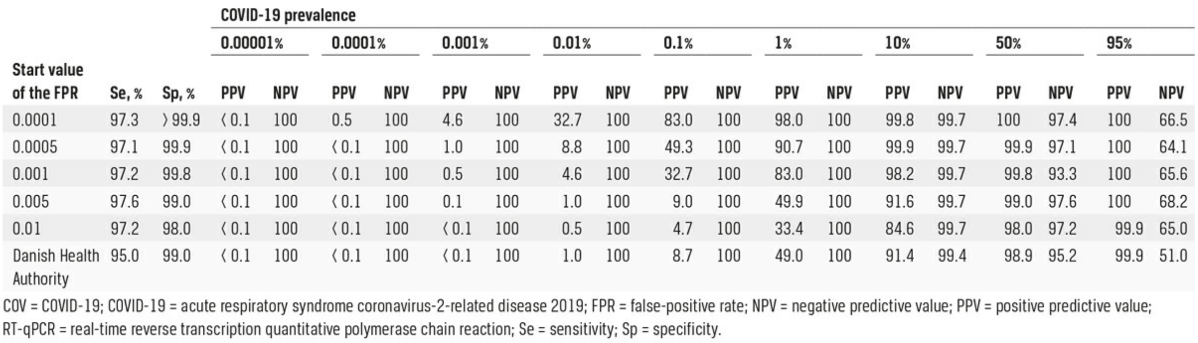
	Unrestricted model		Restricted model	
	COV+	COV–	COV+	COV–
Latent class probability	0.6119	0.3881	0.9018	0.0982
Conditional probability				
+RT-QPCR	0.9994[a]	0.6809	0.9727[a]	0.0002
-RT-QPCR	0.0006	0.3191b	0.0273	0.9998[b]
+CT	0.8643	0.1645	0.6532	0.0469
-CT	0.1357	0.8355	0.3468	0.9531

COV = acute respiratory syndrome coronavirus-2-related disease 2019; RT-qPCR = real-time reverse transcription quantitative polymerase chain reaction.

a) Represents estimated sensitivity for RT-qPCR.

b) Represents estimated specificity for RT-qPCR.

TABLE 3 / Percentaged predictive values as a function of acute respiratory syndrome coronavirus-2-related disease 2019 prevalence and the start value of the false-positive rate for real-time reverse transcription quantitative polymerase chain reaction for the latent class "COV–" in the iteration restriction compared with those derived based on estimates reported by the Danish Health Authority.



			COVID-19 prevalence																	
			0.00001%		0.0001%		0.001%		0.01%		0.1%		1%		10%		50%		95%	
Start value of the FPR	Se, %	Sp, %	PPV	NPV	PPV	NPV	PPV	NPV	PPV	NPV	PPV	NPV	PPV	NPV	PPV	NPV	PPV	NPV	PPV	NPV
0.0001	97.3	〉99.9	〈0.1	100	0.5	100	4.6	100	32.7	100	83.0	100	98.0	100	99.8	99.7	100	97.4	100	66.5
0.0005	97.1	99.9	〈0.1	100	〈0.1	100	1.0	100	8.8	100	49.3	100	90.7	100	99.9	99.7	99.9	97.1	100	64.1
0.001	97.2	99.8	〈0.1	100	〈0.1	100	0.5	100	4.6	100	32.7	100	83.0	100	98.2	99.7	99.8	93.3	100	65.6
0.005	97.6	99.0	〈0.1	100	〈0.1	100	0.1	100	1.0	100	9.0	100	49.9	100	91.6	99.7	99.0	97.6	100	68.2
0.01	97.2	98.0	〈0.1	100	〈0.1	100	〈0.1	100	0.5	100	4.7	100	33.4	100	84.6	99.7	98.0	97.2	99.9	65.0
Danish Health Authority	95.0	99.0	〈0.1	100	〈0.1	100	〈0.1	100	1.0	100	8.7	100	49.0	100	91.4	99.4	98.9	95.2	99.9	51.0

COV = COVID-19; COVID-19 = acute respiratory syndrome coronavirus-2-related disease 2019; FPR = false-positive rate; NPV = negative predictive value; PPV = positive predictive value; RT-qPCR = real-time reverse transcription quantitative polymerase chain reaction; Se = sensitivity; Sp = specificity.

DISCUSSION

Our figures underscore the importance of the expected COVID-19 prevalence of the tested population. Early in the epidemic, primarily individuals with severe symptoms were tested and the prevalence in this population was probably high (> 10%), ensuring a high PPV (> 90% based on our main analysis). With broader population testing, e.g., testing persons with mild or no symptoms, a lower prevalence is expected. The prevalence of active COVID-19 can roughly be estimated at between 0.08% and 0.8% based on the number of new cases per day (approximately 40) times the duration of the infectious state (14 days) times the dark figure (between 8 and 80) divided by population size (5.8 mil). Using these prevalence estimates and the high estimates of sensitivity (95%) and specificity (99.98%), the PPV would range between 80% and 97.5%; i.e., between 1/40 and 1/5 positive tests could be false positive. A potential high risk of false positives needs to be considered by clinicians and decision makers. If the consequence of a positive test result were quarantining, the impact of widespread testing on disease transmission would be limited. Conversely, false-positive individuals interned in multi-bed rooms or halls would be at high risk of becoming infected by truly infected co-interned individuals. A low PPV may also lead to underestimates of the true case-fatality rate, possibly leading to decreased public awareness and adherence to recommendations for reduction of virus transmission.

The risk of false-negative results as a function of prevalence should also be considered. The recommendation for, e.g., healthcare personnel showing COVID-19 symptoms but being outside identified local outbreaks, is quarantining until a negative test result or until 48 hours after symptoms have resolved if testing positive. If the previous prevalence estimate of 0.08% is used, the NPV would be 99.98%. However, if healthcare personnel are tested during a local COVID-19 outbreak (e.g., in a nursing home), the prevalence may be estimated to reach 25% and the NPV would be approximately 90%. By repeating the test three times at two seven-day intervals, the risk of false-negative results would shift from 1/10 to 1/1,000, thus reducing the risk of maintaining the outbreak.

Estimating the prevalence of COVID-19 is also a challenge due to the enormous span in clinical symptoms between infected individuals and the concomitant uncertainty in the estimation of the dark figure; and some may find the prevalence of ≤ 0.8% and perhaps even less than 0.08 in our example above to be an underestimate. However, it should be kept in mind that it is the prevalence (or incidence proportion) of ongoing infection that is relevant for RT-qPCR tests; this is likely to be low in the current context of extended testing and decreasing rates of transmission. The situation is different for antibody tests that will detect infected (in late-stage) and recovered persons (point prevalence) alike; thus, prevalence figures are higher. On 29 May, Copenhagen University Hospital, Hvidovre reported a sensitivity of 93% and a specificity of 98.3% for a SARS-CoV-2 antibody test from Wantai. In a high-prevalence population, this will yield a high PPV, thus making it suitable as a confirmatory ELISA-based test for persons with a positive RT-qPCR test). A limitation is that the antibody test may have to be delayed or repeated if the person presents early in the disease course, as seroconversion does not occur until two weeks after infection. Similar to

RT-qPCR, the use of antibody tests in a low-prevalence setting (e.g., in the general population early in the epidemic) carries a higher risk of false-positives. The consequence could be an overestimation of population immunity with an inadvertent negative impact on effective measures such as social distancing and hygienic measures as well as lower adherence to future vaccine recommendations.

Our estimates from the LCA share the limitations inherent in the study by Ai et al [9]. That study was not specifically designed as a diagnostic test study. Test indications were not reported, but all patients presumably had symptoms in the severe end of the disease spectrum. Furthermore, physicians interpreting the CTs were not blinded to other clinical patient data. Issues with the spectrum effect and subjective errors in categorisation likely led us to overestimate the sensitivity and specificity in the LCA, especially when applied to a broader population [14]. Indeed, we did observe very high values in our restricted model. However, this only secures conservative estimates of the derived predictive values. The spectrum effect may also have led to overestimates of sensitivity and specificity reported by the studies used for our main analysis.

Another limitation of our LCA is the low specificity of RT-qPCR using the initial unrestricted LCA. Unrestricted LCA may give erroneous results when performed on only two observable parameters (e.g., CT and PCR), which may explain the low specificity observed in that analysis. A way of improving the model could have been to include additional clinical information (e.g., symptoms, risk behaviour, other test results), but we had no such data. Instead, we placed a restriction on the false-positive rate. Although the sensitivity and specificity varied only slightly in the different restricted models, it cannot be ignored that the predictive values are dependent on model specification. It also illustrates that predictive values for RT-qPCR testing depend on, e.g., issues with primer purity, test equipment stability and procedure stringency, and such factors may have changed during the epidemic due to demands of accelerated test facilities.

CONCLUSIONS

A high risk of false-positive RT-qPCR tests should be considered when expanding the test strategy, whereas false-negatives may occur during local outbreaks. A confirmatory test (e.g., demonstrating seroconversion or repeated RT-qPCR) may be warranted. LCA may be used to estimate test performance using multiple diagnostic tests when a gold standard is unavailable. Although there are limitations to LCA, the method may be useful for future epidemics, and there is a potential to expand the LCA with further clinical information and new diagnostic tests as they emerge.

CORRESPONDENCE: *Henrik Frank Lorentzen*. E-mail: lorentzen@dadlnet.dk

ACCEPTED: 14 July 2020

CONFLICTS OF INTEREST: none. Disclosure forms provided by the authors are available with the full text of this article at Ugeskriftet.dk/dmj

LITERATURE

1. SSI. COVID-19 i Danmark Epidemiologisk overvågningsrapport 2020 07 April 2020. https://files.ssi.dk/COVID19-overvaagningsrapport-07042020-wvp1 (2 Jun 2020).
2. Tang Y-W, Schmitz JE, Persing DH et al. Laboratory diagnosis of COVID-19: current issues and challenges. J Clin Microbiol 2020;58:e00512-20.
3. Danish Health Authority. Information om PCR test for COVID-19 til almen praksis2020 14.04.2020. www.sst.dk/-/media/Udgivelser/2020/Corona/IRF-Almen-praksis/Kommunikation-til-almen-praksis_test-af-COVID.ashx?la=da&hash=A28AC860410928AAD8864F689B6A05C9E56F3A87 (14 Apr 2020).
4. Wang W, Xu Y, Gao R et al. Detection of SARS-CoV-2 in Different types of clinical specimens. JAMA 2020;323:1843-4.
5. InstitutPasteur. Protocol: real-time RT-PCR assays for the detection of SARS-CoV-2. www.who.int/docs/default-source/coronaviruse/real-time-rt-pcr-assays-for-the-detection-of-sars-cov-2-institut-pasteur-paris.pdf?sfvrsn=3662fcb6_2 (23 Jun 2020).
6. McCutceon AL. Latent class analysis. Delaware: Sage, 1987.
7. Vermunt JK. Latent class models. lEM: a general program for the analysis of categorical data 1. Tilburg: Department of Methodology and Statistics, Tilburg University, 1997.
8. Kingdom FAA, Prins N. Chapter 9 - Model comparisons. In: Kingdom FAA, Prins N, eds. Psychophysics. 2nd ed. San Diego: Academic Press, 2016:247-307.
9. Ai T, Yang Z, Hou H et al. Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology 2020:200642.
10. Uebersax J. LCA frequently asked questions (FAQ) 2006. www.john-uebersax.com/stat/faq.htm (6 Apr 2020).
11. Vermunt JK. win-LEM. 1 ed. Tilburg: Tilburg University, 1996.
12. MedCalc Statistical Software. 19.2.1 ed. Ostend, Belgium: MedCalc Software Ltd, 2020.
13. R_Core_Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2019.
14. Usher-Smith JA, Sharp SJ, Griffin SJ. The spectrum effect in tests for risk prediction, screening, and diagnosis. BMJ 2016;353:i3139.