Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company’s public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
The estimation of diagnostic accuracy of tests for COVID-19: A scoping review

Dierdre B. Axell-House a, Richa Lavingia b,c,d, Megan Rafferty c,d, Eva Clark a,e, E. Susan Amirian e, Elizabeth Y. Chiao *a,e

a Section of Infectious Diseases, Department of Internal Medicine, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA
b Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA
c School of Social Sciences, Rice University, MS 272, 5620 Greenbriar Dr, Houston, TX 77005, USA
d UTHealth School of Public Health, Houston, TX, USA
*e Houston HSRI/ID RxES, Michael E. DeBakey VA Medical Center, 2450 Holcombe Blvd, Houston, TX, 77021, USA
e Department of Epidemiology, The University of Texas MD Anderson Cancer Center, 1555 Pressler St., Unit 1340, Houston, TX 77030, USA

ARTICLE INFO

Article history:
Accepted 27 August 2020
Available online 31 August 2020

Keywords:
SARS-CoV-2
COVID-19
Diagnostic accuracy
Sensitivity
Specificity
QUADAS-2

SUMMARY

Objectives: To assess the methodologies used in the estimation of diagnostic accuracy of SARS-CoV-2 real-time reverse transcription polymerase chain reaction (rRT-PCR) and other nucleic acid amplification tests (NAATs) and to evaluate the quality and reliability of the studies employing those methods.

Methods: We conducted a systematic search of English-language articles published December 31, 2019-June 19, 2020. Studies of any design that performed tests on ≥10 patients and reported or inferred correlation statistics were included. Studies were evaluated using elements of the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) guidelines.

Results: We conducted a narrative and tabular synthesis of studies organized by their reference standard strategy or comparative agreement method, resulting in six categorizations. Critical study details were frequently unreported, including the mechanism for patient/sample selection and researcher blinding to results, which lead to concern for bias.

Conclusions: Current studies estimating test performance characteristics have imperfect study design and statistical methods for the estimation of test performance characteristics of SARS-CoV-2 tests. The included studies employ heterogeneous methods and overall have an increased risk of bias. Employing standardized guidelines for study designs and statistical methods will improve the process for developing and validating rRT-PCR and NAAT for the diagnosis of COVID-19.

© 2020 The British Infection Association. Published by Elsevier Ltd. All rights reserved.

Introduction

After its emergence in December 2019, the virus now known as SARS-CoV-2 was identified and sequenced in early January 2020, allowing for the rapid development of diagnostic testing based on the detection of viral nucleic acid (i.e., real-time reverse transcription polymerase chain reaction [rRT-PCR]). Because infected patients can present with non-specific symptoms or be asymptomatic, the development of accurate diagnostic tests for both clinical and epidemiological purposes was a crucial step in the response to the COVID-19 pandemic.

In the United States, the spread of SARS-CoV-2 rapidly outpaced the capacity to test for it, resulting in the Food and Drug Administration (FDA) relaxing regulatory requirements to increase testing availability. The FDA granted the first Emergency Use Authorization (EUA) for a SARS-CoV-2 rRT-PCR diagnostic test on February 4, 2020. Consequently, hundreds of tests for SARS-CoV-2, among them rRT-PCRs, other types of nucleic acid amplification tests (NAATs), and automated and/or multiplex methods based on proprietary platforms, obtained FDA Emergency Use Authorization (EUA). As of August 4th, 2020, the FDA has granted EUAs to 203 diagnostic tests, including 166 molecular tests, 35 antibody assays, and 2 antigen tests. Although the FDA began requiring the submission of validation methods and results as part of EUA application for SARS-CoV-2 diagnostic tests, these tests were not initially required to undergo the rigorous assessment that would normally be part of the FDA approval process. Researchers also began developing alternative nucleic-acid based methodologies to detect SARS-CoV-2, including reverse-transcription loop-mediated isothermal amplification (RT-LAMP), and others.

* Corresponding author.
E-mail address: eychiao@mdanderson.org (E.Y. Chiao).

https://doi.org/10.1016/j.jinf.2020.08.043
0163-4453/© 2020 The British Infection Association. Published by Elsevier Ltd. All rights reserved.
Concurrently with rapid test production, publications emerged reporting clinical diagnostic test performance characteristics, such as “sensitivity” and “specificity”, though some lacked the rigorous methodologies usually required to formally estimate diagnostic accuracy. Here we present a scoping review of the literature with two main objectives: 1) to assess the methodologies used in the estimation of diagnostic accuracy of SARS-CoV-2 tests and 2) to evaluate the quality and reliability of the studies employing those methods.

Methods

Data sources and searches

Searches were performed through MEDLINE (Ovid), EMBASE (Elsevier), Scopus, Web of Science, CINAHL, and PubMed following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines between December 1, 2019 and June 19, 2020. The following search string was used: (2019-nCoV or SARS-CoV-2 or SARS-CoV2 or COVID-19 or COVID19 or COVID) and (“positive agreement” or “negative agreement” or “overall agreement” or “diagnostic accuracy” or “positive rate” or “positivity rate” or “test performance” or “reference standard” or “gold standard” or sensitivity or specificity or “percent agreement” or “concordance” or “test agreement” or “predictive value” or “false negative” or “false positive”) and (“polymerase chain reaction” or PCR or “reverse transcriptase” or “nucleic acid amplification test” or NAAT or isothermal or “RT-LAMP” or “RT-PCR” or “molecular test”). The literature hub LitCovid’s “Diagnosis” section was screened in its entirety once and then daily for relevant titles.

Study selection

We liberally screened articles by title and abstract for further evaluation. Articles were included if they met the following criteria on screening: 1) Peer-reviewed publication, 2) Study evaluated diagnostic test accuracy of NAAT, 3) Diagnostic test performed on ≥10 patients, 4) Diagnostic/clinical sensitivity, specificity, other correlates, or test positive rate were either identified by name or were included in the publication as a numerical value and we could reproduce the calculations. Exclusion criteria included: 1) Pre-print status, 2) Guidelines, consensus, review, opinion, and other summary articles 3) Entirely pregnant or pediatric populations, 4) Overlap of study population with another included publication.

Data extraction and quality assessment

Four authors independently extracted data and two authors reviewed data for accuracy. For study characteristics, we extracted: first author name, country, study design, patient population, total number of patients or samples included in test performance calculations, and number of cases according to rRT-PCR (Tables 1–5) or total number of cases based on positive result of any platform tested (Table 6). For patient characteristics, we extracted age and sex. For index test and reference standard characteristics, we extracted: test type (NAAT or definition (clinical diagnosis, composite reference standards), specimen (NAAT), specimen dry/collection liquid status (for studies evaluating Abbott ID NOW), proprietary automated and/or multiplex systems – henceforth called “platforms” (NAAT), and target genes of primers (NAAT). For outcomes, we extracted the values of test performance characteristics with their designation according to the original authors, without our interpretation. For this reason, we indicate these outcomes as “reported” (r); reported sensitivity (rSN), specificity (rSP), positive predictive value (rPPV), negative predictive value (rNPV), accuracy
Table 2
Studies reporting test performance characteristics of initial rRT-PCR result compared to result after repeated tests of rRT-PCR as reference standard.

| Authors          | Country | Study Type | Demographics‡ | Specimen Type (No.) | Primers/Platform No. | No. pts with 1st rRT-PCR positive | Total No. pts with rRT-PCR ever positive | Total No. pts in calculations* | Interval (d) between each re-test | No. of tests performed per pt until positive | Interval (d) between initial and positive (final) rRT-PCR | Study Findings | Correlative Statistics (95% CI) | PR  |
|------------------|---------|------------|---------------|---------------------|----------------------|-----------------------------------|------------------------------------------|------------------------------------|----------------------------------|---------------------------------------------|---------------------------------------------|---------------|-------------------------------|-----|
| Bernheim et al.  | China   | Cases only | Age (y) 45.3 ± 15.6 | % Male 50.4% | NPS, OPS, Trach Asp, RAL (nr) | ORF1ab, N rRT-PCR | 90 102 | 102 | nr | 1 test: 90 | >1 test: 12 | nr | Correlative Statistics (95% CI) | 88% |
| Fang et al.      | China   | Cases only | Age (IQR 39–55) 45 | % Male 56.9% | Throat swab (45), sputum (6) | ORF1ab, N rRT-PCR | 36 51 | 51 | ≥1 | 1 test: 36 | 2 tests: 12 | 3 tests: 2 | 4 tests: 1 | 1 test: 10,070 | >1 test: 7335 | 1 - 49 | Lower bound estimate: 57.9% (55.2–60.5) | 71% (56–83) |
| Green et al.     | USA     | Cohort     | Age (y) 53.1 ± 22.3 | % Male 45.5% | NPS, OPS (nr) | RdRp, E Roche cobas 19,195 (70.1%) | 10,070 17,405** | 22,061 | median 8 range (1 – 49) | 1 test: 10,070 | >1 test: 7335** | 1 - 49 | Upper bound estimate: 94.6% (94.2–95.0) | 88% |

(continued on next page)
| Authors           | Country     | Study Type   | Demographics† | Specimen Type No. | Primers/ Platform No. | No. pts with 1st rRT-PCR positive | No. pts with 1st rRT-PCR ever positive | Total No. of pts in calculations‡ | Interval (d) between each re-test | No. of tests performed per pt until positive | Interval (d) between initial and positive (final) rRT-PCR | Study Findings | PR |
|-------------------|-------------|--------------|---------------|-------------------|-----------------------|-------------------------------------|----------------------------------------|---------------------------------------|----------------------------------------|-----------------------------------------------|---------------------------------------------|--------------|----|
| He et al.         | Hong Kong   | Case-Control | Age (y)† 52 (8 - 74) 50% | NPS, OPS, Trach Asp, BAL (nr) | RdRp, S rRT-PCR        | 27                                  | 34                                     | 82 nr                                 | 1 test: 27 > 1 test: 7                | 1–14                                        | 79% (66–93) | 100% (100) | 92% (91–92) | 88.6% |
| Lee et al.        | Singapore   | Cases only   | nr nr         | NPS (70)          | ORF1ab, N rRT-PCR     | 62                                  | 70                                     | 70 nr                                 | 1 test: 62 2nd-Xkb: 1–2                | 1st-2nd: Final:                             | 83.3%                                      | 91% (83–97) | 51.25% |
| Long et al.       | China       | Cases only   | 44.8 ± 18.2   | 55.6%             | OPS, NPS (nr)         | 30                                  | 36                                     | 36 nr                                 | 1 test: 64 1st: 6 2nd: 3                 | 300 test: 6 1st: 2 2nd: Final: 2.47 2 - 8 | 83.3%                                      | 91% (83–97) | 51.25% |
| Wong et al.       | Hong Kong   | Cases only   | 56 ± 19       | 40.6%             | NPS, throat swab (nr) | 58                                  | 64                                     | 64 nr “not uniform”                   | 1 test: 58 > 1 test: 6                  | 6 test: 3 1st: 2 2nd: Final:             | 83.3%                                      | 91% (83–97) | 51.25% |
| Wu et al.         | China       | Cases only   | 46.1 ± 15.4   | 48.8%             | Nose swab, throat swab (nr) | 41                                  | 80                                     | 80 1                                  | 1 test: 41 2 test: 30                   | 3 tests: 9                                  | 83.3%                                      | 91% (83–97) | 51.25% |

* Pts included in test performance calculations.
† Of cohort or cases.
‡ Format: median (range), median(IQR), or mean±SD. Patient Population.
§ Hospitalized patients. Abbreviations- BAL: Bronchoalveolar lavage, CI: confidence interval, d: days, E: envelope, HEL: helicase, IQR: Interquartile range, N: nucleocapsid, No.: number, NPS: nasopharyngeal swab, nr: not reported, OPS: oropharyngeal swab, ORF1ab: open reading frame 1ab, PR: positive rate, pts: patients, rRT-PCR: real-time Reverse Transcription Polymerase Chain Reaction, rAcc: reported accuracy, RdRp: RNA-dependent RNA polymerase, rSN: reported sensitivity, rSP: reported specificity, S: spike, Trach Asp: Tracheal Aspirate, y: years.
** Sensitivity estimates for the first test conducted on patients were calculated based on different assumptions about true negatives. The estimate of the upper bound estimate assumes that any negative test results (whether negative on a single test or consistently negative across multiple, repeated tests) was a true negative (aka, false negative rate=0%). The estimate of the lower bound uses the proportion of repeatedly tested cases who initially tested negative but then tested positive in repeated tests to calculate a false negative rate (16.8%) and apply that rate to the patients who only received a single test to calculate an assumed number of false negative cases. Additional details are provided in Suppl. Fig. 1 and Green et al.
Table 3
Studies that calculate test performance characteristics of rRT-PCR or automated rRT-PCR platforms compared to composite reference standards.

| Authors          | Country | Study Type | No. Patients | Demographics | Specimen | Index Test | Composite Reference Standard Definition | Study Findings (95% CI) |
|------------------|---------|------------|--------------|--------------|----------|------------|------------------------------------------|------------------------|
| Cradic et al.    | USA     | Cohort† 5 | 184          | 33           | NPS in VTM | Automated Multiplex rRT-PCR | Result obtained from at least 2 of the 3 assays is consensus result. | 100% (90–100) 100% (98–100) |
| Suo et al. China | China   | Cohort** 2 | 58           | 52           | Throat swab | Automated Multiplex rRT-PCR (China CDC protocol) | Positive result of repeated rRT-PCR, or serology is considered a positive result. | 40% (27–55) 100% (54–100) 100% (47–57) 16% (13–19) 47% (33–60) |
| Zhen & Mangi     | USA     | Case       | 104          | 51           | NPS       | rRT-PCR (US CDC protocol) | Result obtained by 3 out of 4 assays tested is consensus result. | 100% (93–100) 98% (89–99) 0.98 (0.94–1) |

† Cases according to composite reference standard.
‡ Of cohort or cases.
† Format: median (range), median(IQR), or mean±SD. Patient Population.
‡† Hospitalized patients.
¶ Emergency Room patients.
** Outpatients, some of whom were later hospitalized.
†† Not reported.

Suprotected data is also present in Table 4. Abbreviations- E: envelope, IQR: Interquartile range, κ: kappa coefficient, M: male, N/A: not applicable, N: nucleocapsid, NPS: nasopharyngeal swab, nr: not reported, ORF1ab: open reading frame 1ab, rRT-PCR: real-time Reverse Transcription Polymerase Chain Reaction, rAcc: reported accuracy, RdRp: RNA-dependent RNA polymerase, rNP: reported negative percent agreement, rPPA: reported positive percent agreement, rPPV: reported positive predicted value, rSN: reported sensitivity, rSP: reported specificity, S: spike, VTM: viral transport media, y: years.
Table 4
Studies reporting test performance characteristics of other nucleic acid amplification test methods compared to rRT-PCR.

| Authors            | Country | Study Type | No. Patients | Cases | Demographics  | Specimen | Index Test Type | Primers | Platform | Ref Std: rRT-PCR | Study Findings (95% CI or p-value) | Cohen’s 𝜂 |
|--------------------|---------|------------|--------------|-------|---------------|----------|-----------------|----------|----------|-----------------|------------------------------------|-----------|
| Baek et al. 21     | Korea   | Case Control† | 154          | 14    | nr            | Nasal swab (154) | RT-LAMP N | nr     | ORF1ab, S     | 100%                                | 98.7%  | 0.826 |
| Kitagawa et al. 22 | Japan   | Cohort†     | 76           | 30    | nr            | NPS (76)  | RT-LAMP nr | ORF1ab, N     | 100%                                | 95.6%  | 97.4% |
| Lau et al. 21      | Malaysia| Case Control† | 89††         | 47††  | nr            | NPS (89)  | RT-LAMP N | RdRp,E         | 100%                                | 100%   | 92.9% |
| Lu et al. 24       | China   | Case Control† | 56           | 36    | nr            | Throat swab (56) | RT-LAMP N | ORF1ab, S     | 100%                                | 97.6%  | 95.2%  |
| Yan et al. 25      | China   | Cohort†     | 130          | 58    | nr            | Throat swab, BAL (nr) | RT-LAMP ORF1ab, S | ORF1ab, N | 100%                                | (92.3–100) | 100%   | (93.7–100) | 96.2%  | (93.4–97.8) | 98.6%  | (97.3–99.1) | 0.952 | p < 0.001 |
| Wang, Cai, & He et al. 26 | China | Cohort† | 947          | 338   | 44 ± 17.1 | Throat swab, sputum, BAL, blood (nr) | RT-RAA ORF1ab | LifeReady 1600 fluores- | ORF1ab, N | 100%                                | 100%   | 1.0  | p < 0.001 |
| Xue et al. 27      | China   | Cohort†     | 120††        | 22††  | nr            | NPS, sputum (nr) | RT-RAA ORF1ab | LifeReady 1600 fluores- | ORF1ab, S | 100%                                | 100%   | 1.0  | p < 0.001 |
| Perchetti et al. 28 | USA     | Case Control† | 356          | 186   | nr            | NPS (356) | Triplex rRT-PCR N1, N2 | ORF1ab, N | 100%                                | 98.4%  | 99.2% |
| Waggoner et al. 29 | USA     | Cohort†     | 27           | 11    | nr            | NPS, OPS (nr) | Triplex rRT-PCR N2, E | ORF1ab, N | 100%                                | 100%   | 100% |
| Li et al. 30       | China   | Cohort†     | 303††        | 126†† | nr            | Throat swab (267), sputum (22), nose swab (8), BAL (3), blood (3) | AIGS ORF1ab, N, S | LifeReady 1000 fluores- | ORF1ab, N | 97.62%                                | (93.2–99.5) | 100% |

(continued on next page)
| Authors          | Country | Study Type | No. Patients Total | Cases* | Demographics† | Specimen (No.) | Index Test Type | Primers | Platform     | Ref Stnd: rRT-PCR Primers | Study Findings (95% CI or p-value) | Cohen’s κ |
|------------------|---------|------------|--------------------|--------|---------------|----------------|----------------|----------|--------------|---------------------------|----------------------------------|----------|
| Suo et al.       | China   | Cohort     | 58                 | 52     | nr            | Throat swab (58) | ddPCR          | ORF1ab, N | QX200 System | ORF1ab, N                  | 94% (83–99)                      | 0.85     |
| Bulterys et al.  | USA     | Cohort     | 80                 | 30     | nr            | NPS (80)       | ORF1ab, N      | Atila iAMP kit | E             | 82.8% (65.0–92.9)            | 0.86 (0.74–0.98)                 | 0.737    |
| Wang, Cai, & Zhang et al. | China | Cohort | 181          | 25     | nr            | Throat swab (181) | OSN-qRT-PCR    | ORF1ab, N | Life Tech. 480 | ORF1ab, N                  | 63% (36–83)                      |          |

* Cases according to reference standard.
† Of cohort or cases.
‡ Format: median (range), median (IQR), or mean ± SD. Patient Population.
¶ Hospitalized patients or Emergency Room/Immediate Care Center patients.
** Outpatients, who were later hospitalized.
†† Number of samples (when number of patients not reported).

4 Suo et al. data is also present in Table 2. Abbreviations- AIGS: Automatic integrated gene detection system, BAL: Bronchoalveolar Lavage, CI: confidence interval, ddPCR: digital droplet polymerase chain reaction, E: envelope, iAMP: isothermal amplification, IQR: Interquartile range, κ: kappa statistic, n/a: not applicable, N: nucleocapsid, No.: number, NPS: nasopharyngeal swab, nr: not reported, OPS: oropharyngeal swab, ORF1ab: open reading frame 1ab, OSN-qRT-PCR: one-step single-tube nested quantitative real-time polymerase chain reaction, rAcc: reported accuracy, rOA: reported negative percent agreement, rNPV: reported negative predictive value, rOA: reported overall agreement, rPPA: reported positive percent agreement, rPPV: reported positive predictive value, rRT-PCR: real-time Reverse Transcription Polymerase Chain Reaction, rSN: reported sensitivity, rSP: reported specificity, RT-LAMP: reverse transcription loop-mediated isothermal amplification, RT-RAA: reverse-transcription recombinase-aided amplification, S: spike, y: years.
Table 5
Studies estimating NAAT platform test performance characteristics compared to rRT-PCR as the reference standard.

| Country | Study Type | No. Patients | Demographics | Specimen | Index Test | Ref Stnd: rRT-PCR Platforms | Study Findings (95% CI) | Cohen’s κ |
|---------|------------|--------------|--------------|----------|------------|----------------------------|------------------------|-----------|
|         |            | Total | Cases | Age (y) | % M | Type | Primers | Platform | rSN | rSP | rPPV | rNPV | rPPA | rNPA | rOA |         |
| Mitchell et al. 31 | USA | Case Control† | 61 | 46 | nr | nr | NPS in VTM (61) | Automated Isothermal NAAT | Abbott ID NOW | N1, N2 | 71.70% | 100% | 94% | 78.70% |
| Rhoads et al. 35,b | US | Cases only† | 96 | 96 | nr | nr | NPS (85), nasal swab (11) | Automated Isothermal NAAT | Abbott ID NOW | N1, N2 | 93% | 100% |          |
| Moore et al. 34,b | USA | Cohort† | 200 | 119 | 50 ± 17 | 46% | NPS in VTM (200) | Automated Isothermal NAAT | Abbott ID NOW | N1, N2 | 96% | 100% |          |
| DegliAngeli et al. 36 | USA | Case Control† | 60 | 30 | nr | nr | Nasal swab, NPS (nr) in VTM (85) | Automated Isothermal NAAT | Abbott ID NOW | N1, N2 | 92% | 100% | 98.4% | 0.97 |
| Hou et al. 37 | China | Cohort† | 285 | 153 | <65 y: 77.2% | 55.8% | NPS (nr) in VTM (285) | Automated Isothermal NAAT | Abbott ID NOW | N1, N2 | 96% | 96.2% | 100% |
| Lieberman et al. 38 | USA | Cohort† | 26 | 13 | nr | nr | NPS (26) | Automated Isothermal NAAT | Abbott ID NOW | N1, N2 | 95.5% | 95.8% | 97% |
| USA, UK, Loeffelholz FR, IT et al. 39,a | USA | Cohort† (enriched for positive cases) | 88 | 13 | nr | nr | NPS (339), NPS+OPS (87), Trach, Asp (30), OPS (15) | Automated Isothermal NAAT | Abbott ID NOW | N1, N2 | 96% | 96.2% | 100% | 0.97 |
| 129 | 60 | S, E | 100% | 100% | (93–100) | 92% | 100% | (94.2–100) |
| 99 | 74 | N1, N2 | 100% | 92% | 100% | (94.2–100) | 97.8% | 100% | (97.8–100) |
| 65 | 30 | RdRp | 100% | 92% | 100% | (88.7–100) | 92% | 100% | (92.0–100) |
| 79 | 35 | RdRp, E, N | 100% | 92% | 100% | (67.6–100) | 92% | 100% | (92.0–100) |
| Bordi et al. 40 | Italy | Cohort+Controls† | 278 | 99 | nr | nr | Nasal swab, NPS (nr) in VTM (99) | Automated Isothermal NAAT | Abbott ID NOW | N1, N2 | 96% | 96.2% | 100% | 0.938 |
| Rhoads et al. 35,b | US | Cases only† | 96 | 96 | nr | nr | NPS (85), nasal swab (11) | Automated Isothermal NAAT | Abbott ID NOW | N1, N2 | 96% | 96.2% | 100% | 0.98 |

(continued on next page)
| Country        | Study Type | No. Patients | Demographics | Specimen Type | Index Test | Ref Std: rRT-PCR Primers | Study Findings (95% CI) | Cohen’s κ |
|---------------|------------|--------------|--------------|---------------|------------|-------------------------|-------------------------|-----------|
| Slovenia      | Cohort (5) | 501          | Cases       | NPS, NPS+OPS | Automated | RdRp, E                 | 100% (92.8–99.9)          | 0.98      |
| USA           | Cohort (5) | 963          | nr          | NPS          | Automated | N1, N2, N3             | 94.2% (98.2–99.9)          | 0.94      |
| Australia     | Cohort (5) | 52           | 31.5 (0–84) | NPS+OPS      | Automated | RdRp, E                 | 100% (92.8–99.9)          | 0.94      |
| USA           | Cohort (5) | 180          | nr          | NPS          | Automated | RdRp, E                 | 100% (92.8–99.9)          | 0.94      |
| Hong Kong     | Cohort (5) | 214          | 91 (IQR 31–69) | NPS          | Automated | RdRp, E                 | 100% (92.8–99.9)          | 0.94      |
| USA           | Case Control (5) | 100          | 50          | NPS          | Automated | RdRp, E                 | 100% (92.8–99.9)          | 0.94      |
| France        | Case Control (5) | 69           | 40          | NPS          | Automated | RdRp, E                 | 100% (92.8–99.9)          | 0.94      |

* Cases according to reference standard.
† Of cohort or cases.
‡ Format: median (range), median(IQR), or mean±SD. Patient Population.
§ Hospitalized patients.
¶ Emergency Room/Immediate Care Center patients.
** Outpatients.
†† not reported.
†‡ Number of samples (when number of patients not reported).
‡‡ Reported as concordance.
§§ Reported as kappa statistic.
¶¶ LFA: lateral flow assay, M: male, n/a: not applicable, N: nucleocapsid, NAAT: nucleic acid amplification test, N Asp: nasopharyngeal aspirate, No.: number, NPS: nasopharyngeal swab, nr: not reported, NS: normal saline, OPS: oropharyngeal swab, ORF1ab: open reading frame 1ab, RdRp: RNA-dependent RNA polymerase, Ref Std: reference standard, rNPA: reported negative percent agreement, rNPP: reported negative predictive value, rOA: reported overall agreement, rPPA: reported positive percent agreement, rPPV: reported positive predictive value, rRT-PCR: real-time Reverse Transcription Polymerase Chain Reaction, rSN: reported sensitivity, rSP: reported specificity, S: spike, Trach Asp: Tracheal Aspirate, UTM: Universal transport medium, VTM: viral transport medium, y: years.
Table 6
Studies assessing agreement between NAAT platforms.

| Authors                     | Country | Study Type | No. Patients | Cases | Demographics | Specimen | Platform #1 | Demographics | Platform #2 | Demographics | Study Findings (95% CI) | rPPA | rNPV | Cohen’s k |
|-----------------------------|---------|------------|--------------|-------|--------------|----------|-------------|--------------|-------------|--------------|------------------------|------|------|-----------|
| Harrington et al. 48        | USA     | Cohort     | 524          | 188   | nr           | Paired NPS in VTM (ReallTime) & foam nasal swab (ID NOW) (524 pairs) | Automated Multiplex RT-PCR RdRp, N | Abbott RealTime | Automated Isothermal NAAT RdRp | Abbott ID NOW | 75% (67.7–80.6) | 99% (97.6–99.8) | 0.87 (0.78–0.96) |
| Moore et al. 49, 50         | USA     | Cohort     | 200          | 125   | 50 ± 17 46%  | NPS in VTM (200) | Automated Multiplex RT-PCR RdRp, N | Abbott RealTime | Automated Isothermal NAAT RdRp | Abbott ID NOW | 75.2% (66.7–82.5) | 100% (95.4–100) | 0.87 (0.78–0.96) |
| Basu et al. 49              | USA     | Cohort     | 101          | 32    | (28 - 90) nr | NPS dry (101) | Automated Multiplex RT-PCR RdRp, N | Abbott RealTime | Automated Isothermal NAAT RdRp | Abbott ID NOW | 54.8% (37.8–70.8) | 83.1% (74.3–99) | 85.1% (76.9–90.8) | 94.4% (73–89.7) |
| Cases only                 | USA     | Case Control | 15           | 15    | nr           | NPS in VTM (15) | Automated Multiplex RT-PCR RdRp, N | Abbott RealTime | Automated Isothermal NAAT RdRp | Abbott ID NOW | 80.4% (66.9–90.2) | 95.9% (86–99.5) | 87.7% (93–100) | 0.922 (0.83–0.99) |
| Hogan et al. JCV, 5–1 50    | USA     | Cohort     | 100          | 53    | nr           | NPS (100) | Automated Multiplex RT-PCR RdRp, N | Abbott RealTime | Automated Isothermal NAAT RdRp | Abbott ID NOW | 80.4% (66.9–90.2) | 95.9% (86–99.5) | 87.7% (93–100) | 0.922 (0.83–0.99) |
| Zhen, Smith et al. 51, b   | USA     | Cohort     | 108          | 58    | nr           | NPS in VTM or UTM (108) | Automated Multiplex RT-PCR RdRp, N | Abbott RealTime | Automated Isothermal NAAT RdRp | Abbott ID NOW | 73.9% (63.2–82.3) | 100% (83.4–100) | 0.922 (0.83–0.99) |
| Smithgall et al. 52         | USA     | Case Control | 113         | 90    | 65 (0–101) 60.2% | NPS in VTM or UTM (113) | Automated Multiplex RT-PCR RdRp, N | Abbott RealTime | Automated Isothermal NAAT RdRp | Abbott ID NOW | 73.9% (63.2–82.3) | 100% (83.4–100) | 0.922 (0.83–0.99) |
| Moran et al. 53             | USA     | Cohort     | 103          | 42    | nr           | NPS (95) | Automated Multiplex RT-PCR RdRp, N | Abbott RealTime | Automated Isothermal NAAT RdRp | Abbott ID NOW | 98.9% (92.9–100) | 92% (72.4–98.6) | 99% |
| Craney et al. 54            | USA     | Cohort     | 389          | 147   | nr           | NPS (389) | Automated Multiplex RT-PCR RdRp, N | Abbott RealTime | Automated Isothermal NAAT RdRp | Abbott ID NOW | 98.9% (92.9–100) | 92% (72.4–98.6) | 99% |
| Zhen, Smith et al. 51, b   | USA     | Cohort     | 108          | 58    | nr           | NPS in VTM (108) | Automated Multiplex RT-PCR RdRp, N | Abbott RealTime | Automated Isothermal NAAT RdRp | Abbott ID NOW | 98.9% (92.9–100) | 92% (72.4–98.6) | 99% |

(continued on next page)
Table 6 (continued)

| Authors                     | Country       | Study Type            | No. Patients | Patients Demographics$^i$ | Specimen Type | Platform #1     | Platform #2     | Study Findings (95% CI) | Cohen's $\kappa$ |
|-----------------------------|---------------|-----------------------|--------------|----------------------------|---------------|-----------------|-----------------|--------------------------|------------------|
| Loeffelholz et al.$^{19,4}$ | USA, UK, FR, IT | Cohort$^{II}$ (enriched for positive cases) | 18$^{II}$    | 8$^{II}$                   | NPS, OPS, Automated Multiplex RdRp, N Trach Asp (nr) | Abbott RealTime | Automated Multiplex RT-PCR | 100% (67.6–100) | 100% (77.2–100)               |
| Norz et al.$^{55}$          | Germany       | Case Control$^{II}$   | 165$^{II}$   | 36$^{II}$                  | NPS, OPS (nr) Automated Multiplex RT-PCR ORF1, E Roche cobas 6800 Automated E RT-PCR | NeuMoDx 96     | 100% | 99.2%               |

$^*$ Case estimated as a positive result of any evaluated platform.
$^i$ Of cohort or cases.
$^{II}$ Format: median (range), median(IQR), or mean±SD. Patient Population.
$^I$ Hospitalized patients.
$^†$ Emergency Room/Immediate Care Center patients.
$^\ast$ Outpatient.
$^\ast\ast$ Not reported.
$^{II}$ Number of samples (when number of patients not reported).
$^a$ Loeffelholz et al. and Moore et al. also appear in Table 5.
$^b$ Zhen, Smith et al. appears twice in Table 6 for ease of comparison of studies of the same platform.
$^c$ Dry or VTM status not reported. Rhoads et al. appears twice in Table 5 for ease of comparison of studies of the same platform.

Abbreviations: CI: confidence interval, E: envelope, IQR: Interquartile range, $\kappa$: kappa statistic, M: male, N: nucleocapsid, NAAAT: nucleic acid amplification test, No.: number, NPS: nasopharyngeal swab, nr: not reported, OPS: oropharyngeal swab, ORF1ab: open reading frame 1ab, RdRp: RNA-dependent RNA polymerase, rNPA: reported negative percent agreement, rNPV: reported negative predictive value, rOA: reported overall agreement, rPPA: reported positive percent agreement, rPPV: reported positive predictive value, rRT-PCR: real-time Reverse Transcription Polymerase Chain Reaction, S: spike, Trach Asp: Tracheal Aspirate, UTM: universal transport medium, VTM: viral transport medium, y: years.
(rAcc), positive percent agreement (rPPA), negative percent agreement (rNPA), overall agreement (rOA), and Kappa coefficient. Additionally, we extracted “positive rate,” a non-standard term used by the included studies to refer to the number of positive NAATs in a population of patients suspected to have COVID-19 (Table 1), or to the number of positive samples in a total population of positive samples after repeat testing (Table 2). We constructed 2 × 2 contingency tables and reproduced test performance characteristic calculations to demonstrate the methods of how the original authors obtained the values (Supplementary Table 1). We report additional pertinent study data in Supplementary Table 2: enrollment dates, number of sites of enrollment, symptomatic status, and chest radiology status. No articles were excluded on the basis of quality in order to present the most comprehensive summary of the currently available evidence.

Data synthesis and analysis

We presented the extracted data in tabular form mirrored by a descriptive synthesis in two broad categories: diagnostic accuracy studies for rRT-PCR (Tables 1–3), and diagnostic accuracy or comparative agreement studies of two NAATs (Tables 4–6). Tables are thematically divided based on the reference standard strategy, or approach to obtaining comparative agreement measures. Diagnostic accuracy studies for rRT-PCR were arranged alphabetically in tables by first author last name (Tables 1–3). Diagnostic accuracy and comparative agreement studies for two NAATs were arranged by decreasing order of studies per methodology, then alphabetically by methodology or platform (Tables 4–6) for easy comparison. Due to significant diversity in methods and reporting of results, we reported grouped summary data for study characteristics, patient characteristics, and outcomes.

We used the framework of the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) to evaluate our selected articles (Supplementary Table 3). We collected data, or noted their absence, for a narrative description of risk of bias and concerns of applicability based on the QUADAS domains. For assessment of bias in patient selection, we evaluated author conflicts of interest, study design type, inclusion/exclusion criteria, method of patient enrollment, and reporting of patient demographics and characteristics (i.e. symptomatic status). For assessment of bias in reference standard and index test, we evaluated the accuracy of the reference standard, the description of duration of symptoms at time of testing, whether the threshold to determine a positive test was pre-specified, and researcher blinding to reference standard and index test results. For assessment of bias in flow and timing, we evaluated whether the reference standard was the same for all patients, the sequence and timing of the performance of the reference standard and index test, whether test performance characteristics were calculated based on sample numbers or patient numbers, and whether indeterminate or invalid results were included in test performance calculations.

Results

Study selection

Our search yielded 1537 articles, with 816 unique articles after deduplication. After screening title and abstract, 130 articles underwent full text evaluation. Ultimately, 49 articles were included in our review (Fig. 1).

The performance of rRT-PCR compared to case definitions or clinical diagnoses

Three studies, with 19 to 1014 patients, report a “positive rate” as the number of positive rRT-PCR out of the number of suspected cases of COVID-19, with a range of 38.42% to 59% (Table 1).7–9 The studies do not report these values as “sensitivity” directly; however, these values were interpreted as reflective of the accuracy of rRT-PCR. Ai et al. compared the positive rate of rRT-PCR (59%) with the positive rate of Chest CT in order to draw conclusions about the accuracy of Chest CT for the diagnosis of COVID-19,7 and Liu et al. and Xie et al. expressed concern that their low calculated positive rates (38.42% and 47.4%, respectively) were indicative of a failure of rRT-PCR to diagnose COVID-19.7,9

In terms of quality assessment, the studies lack specific details as to how patients were classified as having suspected COVID-19 infection. The accuracy of clinical diagnosis based on case definitions is unclear but is likely not ideal for diagnosis. Additionally, duration of symptoms at the time of clinical diagnosis or rRT-PCR testing was not provided (Supplementary Table 3).

The performance of rRT-PCR compared to end result after multiple repetitions of rRT-PCR

Eight studies, with a range of 36 to 22,061 patients per study, attempted to determine the accuracy of rRT-PCR by comparing the initial rRT-PCR result to the result after multiple repeated samples from the patient submitted for rRT-PCR testing, which was called the reference standard (Table 2).10–17 Three studies reported this value as a “positive rate,” ranging from 51.25% to 88%,10,14,17 and five reported sensitivity, with a range of 57.5% to 94.6%.11–13,15,16 Of these studies, only He et al. included an rSN of 100%, calculated from patients who remained negative for SARS-CoV-2 after repeated sample testing (Supplementary Table 1).13

Green et al. included patients in their study regardless of whether they were tested once or multiple times, using data from these subsets of patients to make assumptions for estimating clinical test characteristics. In addition, this study also conducted multiple different NAATs and rRT-PCRs on patients, whereas other studies employing this strategy used only one type of NAAT. Additionally, the authors do not clarify whether patients who had repeat SARS-CoV-2 test were consistently tested with the same NAAT/rRT-PCR test or a different one. They also calculated test performance characteristics differently from other studies: two estimates of sensitivity were calculated, one in which the rate of false negatives for single-tested patients was 0%, and one in which the “false negative” rate was the same as in repeat-tested patients in their study of approximately 16.8%.12 However, the details of how they calculated test characteristics were not presented. To clarify the two assumptions made in the calculations, we reconstructed the calculation in Supplementary Figure 1 which demonstrated the range of rSN with an estimate of the lower bound of 57.9% (55.2%–60.5%) and an estimate of the upper bound of 94.6% (94.2%–95.0%).

In terms of quality assessment, most of the studies were performed with non-cohort design, and six consisted of only patients who were determined to have COVID-19 by rRT-PCR, i.e. cases only (Table 2).10,11,13–16 Five of the studies had inclusion criteria which caused pertinent patients to be excluded by necessitating patients to have had a well-performed CT Chest or X-Ray (Supplementary Table 3). This excluded several patients who would otherwise have been pertinent to the study of test diagnostic accuracy.10,11,13,16,17 The studies involved repeating rRT-PCR several times for a reference standard, but each patient received a different number of repeat tests over a different time period, resulting in each patient receiving a different reference standard. One study tracked negative-
to-positive conversion over 1 to 49 days, and another tracked over 1 to 14 days, leading to concern that potentially a patient could have been infected in the time between the initial test and the final test and confounding results. One study counted invalid results as negative results and indeterminate results as positive results when calculating test performance characteristics, otherwise the rationale and ways invalid and indeterminate results were handled were not reported in these studies.

The performance of rRT-PCR compared to various composite reference standards

Three studies determined the accuracy or agreement of rRT-PCR or automated rRT-PCR platforms/instruments compared to a reference standard based on the results of several tests as a “composite reference standard” (Table 3). There were between 58 and 184 patients per study. Suo et al. considered a positive result of either repeated measurements of rRT-PCR or serology to indicate a positive test according to the reference standard; reported sensitivity of initial rRT-PCR result was 40%, rSP 100%, rPPV 100%, and rNPV 16%. Zhen et al. compared rRT-PCR performed according to the US CDC protocol to a composite reference standard in which the consensus result of 3 or more out of 4 molecular assays was considered the correct result. The rRT-PCR had an rPPA of 100%, an rNPA of 99%, and Cohen’s kappa coefficient of 0.98. Cradic et al. did not study rRT-PCR but studied three automated molecular assays and used a composite reference standard of the consensus result of two or more of the three assays. While Abbott ID NOW had a rPPA of 91%, the Roche cobas 6800 and Diasorin Simplexa assays had a rPPA of 100%.

These studies either did not report how samples were selected for evaluation (Supplementary Table 3), or reported that only samples which had sufficient residual volume and had been properly stored were selected. Cradic et al. and Zhen & Mangi et al. had initially tested samples with one platform, and some or all samples were frozen, and subsequently thawed and tested with other platforms, leading to confounding factors in the reference standard and test performance calculations involving the various platform results.

Suo et al. used repeat rRT-PCR testing as part of the reference standard, with repeat tests performed 2–10 days after the initial test, after the patient had been discharged from the hospital, leading to potential exposure for initial infection or reinfection.

The performance other nucleic acid amplification test methods compared to standard rRT-PCR

Fourteen studies compared other nucleic acid amplification test methods to detect SARS-CoV-2 to rRT-PCR (Table 4), with between 27 and 356 patients per study. Five of the studies evaluated reverse transcription loop mediated isothermal amplification (RT-LAMP), four reported sensitivity of 100% and specificity of 95.6 to 100%, and one study reported accuracy of 92.9%. Two studies, Wang, Cai, & He et al. and Xue et al., evaluated reverse-transcription recombinase-aided amplification (RT-RAA) with Cohen’s kappa of 0.952 and 1.0. Two studies, Perchetti et al. and Waggoner et al., evaluated triplex rRT-PCR, reporting overall agreement as 99.2 and 100%, and Li et al. evaluated an automatic integrated gene detection system (AIGS) with rSN 97.2% and rSP 98.5%. Suo et al. evaluated digital droplet polymerase chain reaction (ddPCR), with rSN 94%, rSP 100%, rPPV 100%, rNPV of 63%, and rAcc 95%. Bultery et al. study evaluated an isothermal amplification method with rSN 82.8% and Cohen’s kappa 0.86. Wang, Cai, and Zhang et al. evaluated one-step single-tube nested quantitative polymerase chain reaction (OSN-qRT-PCR) with Cohen’s kappa of 0.737.

Regarding evaluation of quality (Supplementary Table 3), the majority of studies did not report how patient samples were
selected for evaluation. In the study conducted by Bulteys et al., sample selection was a convenience selection of samples with residual volume that had been stored correctly. Most studies did not report symptomatic status of the patient or patient demographics. Problematically, many of the studies did not report when the reference standard was conducted on the patient samples compared to the index test, or whether actions that could potentially alter test results (such as freeze/thaw cycles) occurred between reference standard or index tests. Four studies calculated test performance characteristics based on number of samples rather than number of patients. The management of indeterminate and invalid test results went largely unreported.

The performance of NAAT platforms compared to rRT-PCR as the reference standard

Fifteen studies compared automated NAAT platforms to various rRT-PCR assays to determine test performance characteristics (Table 5), with between 26 and 963 patients or samples per study. Three studies evaluated Abbott ID NOW, an isothermal NAAT platform, with rPNA or rSN of 71.7% to 94%, and rNPA or rSP of 100%. Two studies evaluated Abbott RealTime with rSN or rPPA of 93% to 100%, and rSP or rNPA of 92.4% to 100%. Three studies evaluated Cepheid Xpert Xpress, with rPPA 96.1% to 100%, rNPA 74.3% to 100%, rOA 96.1% to 100%, and Cohen’s Kappa of 0.92. Two studies evaluated Diasorin Simplexa with rSN or rPPA of 96% to 100%, and rNPA of 100%. Two studies evaluated Roche cobas 6800 with rPPA 94.2% to 100%, rNPA 99.5% to 99.6%, and Cohen’s Kappa of 0.90 to 0.98. Other studies evaluated AusDiagnostixs (rSN 100%, rSP 92%),

With regards to quality evaluation (Supplementary Table 3), most studies did not report method of sample collection/patient recruitment and four studies conducted a convenience selection of samples, including enrichment for positive samples. Eight studies conducted test performance calculations on four samples instead of patient numbers. Four studies conducted calculation of test performance characteristics with indeterminate or inconclusive results as “positive,” and the management of indeterminate/inconclusive as well as invalid results went unreported in an additional three studies.

The agreement of NAAT platforms compared to other NAAT platforms

Ten studies, containing between 15 and 524 patients per study, evaluated the agreement between two different types of NAAT platforms (Table 6), typically under the circumstances where one platform was the standard of care at the institution, and another was introduced. The Abbott ID NOW platform, using isothermal amplification, was the most frequently studied test, with an rPPA of 75–75.2% compared to Abbott RealTime, 54.8% compared to Cepheid Xpert Xpress, 80.4–87.7% compared to Hologic Panther Fusion, and 73.9% compared to Roche cobas 6800. Two studies evaluated Cepheid Xpert Xpress compared to Roche cobas 6800, with rPPA 99.8% and rNPA 92% in one, and overall agreement of 99% in another. Several platforms were compared to Hologic Panther Fusion, including cobas 6800 with rOA 96.4%, Cepheid Xpert Xpress with rPPA 98.3% and rNPA 100%, and GenMark ePlex with rPPA 91.4% and rNPA 100%.

In the studies, some platforms were identified as the “comparator” or “reference” platforms, including Cepheid Xpert Xpress, Abbott RealTime, Hologic Panther Fusion, and Roche cobas 6800, and these were listed as “Platform #1” in Table 6. Three studies did not identify any studied platform as the “comparator” or “reference standard,” and instead only reported general, non-directional measures of agreement such as overall agreement, Cohen’s Kappa, or alternatively, the calculations of PPA and NPA were identical no matter their method of calculation (Supplementary Table 1).

Regarding quality evaluation (Supplementary Table 3), the samples used for calculating test performance characteristics were reported to be selected for enrichment of positive samples, diversity of viral load, otherwise curated, or the method of selecting samples was unreported. Symptomatic status of the patients was largely unreported. Five studies included samples where one test was conducted, then interim freezing, cooling, or other storage, before performance of the second test. Two studies did not report the sequence of testing of the two platforms or interim handling or storage of the samples. The status of researcher blinding to either platform result was not reported in any study.

Discussion

In our scoping review of 49 articles concerning test performance characteristics of rRT-PCR and other NAAT used for the diagnosis of COVID-19, we were able to observe several overarching themes. Clinical diagnosis by the case definition for COVID-19 used in the early period of the pandemic does not correlate well with positive rates of COVID-19 rRT-PCR (Table 1). The result of the initial rRT-PCR performed on a patient, if negative, may not be reflective of the result after multiple repeated rRT-PCRs for that patient (Table 2). Several alternative NAAT methods, many of which are easier or faster to perform, may be comparable to standard rRT-PCR (Table 4).

These findings should be viewed cautiously as the SARS-CoV-2 tests in these studies have not undergone rigorous evaluation necessary for FDA approval due to the emergency state generated by the COVID-19 pandemic. In addition, during our scoping review, we found substantial heterogeneity among available studies in terms of test types, reference standards, metrics, and details of study design and methodology.

We categorized the included studies by four different reference standard strategies: clinical diagnosis/case definitions (Table 1), repeated index testing (Table 2), composite reference standard (Table 3), and rRT-PCR (Table 4 and 5). Additionally, we identified a fifth category, where instead of using a reference standard, comparative agreement between two NAAT platforms was calculated (Table 5 and 6).

The main limitation of the first group of studies (Table 1) was the use of a “case definition” as the reference standard to report a “positive rate” of rRT-PCR. During novel disease outbreaks, standard case definitions are often developed to assist clinicians in case identification before a diagnostic test is available. Unfortunately, the studies included in this group were unable to use a clear case definition; instead they refer to a population of “suspected cases,” for which the definition is not reported. Because this group enrolled patients prior to February 15, 2020 in China, during the time in which the Chinese National Guideline for Diagnosis and Treatment of COVID-19 (NGDTC) published five different versions of the COVID-19 case definition, the case definitions in use at the time of these studies varied. A recent study estimated that if a single guideline (specifically, version 5 of the NGDTC) had been used to
identify cases from the beginning of the outbreak to February 20, 2020, there would have been more than three times as many identified cases in Hubei province. This is relevant to our review because the two largest studies that evaluated the rRT-PCR positive rate of patients with a clinical diagnosis of COVID-19 took place in Wuhan, Hubei province, and included patients evaluated before February 14, 2020. \(^7,8\) (Supplementary Table 2). This increased case estimate due to diagnosis of COVID-19 based on case definition complicates the legitimacy and reported accuracy of the “positive rate” of rRT-PCR referred to in these studies.

The second group assessed rRT-PCR test performance characteristics via repeated index rRT-PCR testing (Table 2). Most studies in this group reported “sensitivity” by dividing the number of participants with positive baseline rRT-PCRs by the total number of participants who eventually had a positive rRT-PCR after repeated measurements. While such an approach may have some advantages over the use of a case definition alone as a reference standard, this strategy is, nonetheless, an imperfect solution with its own set of inherent limitations. SARS-CoV-2 infection is transient and the associated viral loads are time-varying because of the natural pathophysiology of the infection. Therefore, the time interval between each repeated test becomes crucially important, and even relatively small time differences (and/or lack of uniformly used intervals) could compromise the interpretation of re-test results and their quality as reference standards. Furthermore, repeated use of the same test as a reference standard for itself does not eliminate the inaccuracies or limitations of the test. Such comparisons ultimately reflect the reliability of the test (assuming a short, uniform time interval between tests), rather than providing a true view of test accuracy.

The third group of three studies calculated test performance characteristics of rRT-PCR according to a composite reference standard (Table 3). Using arbitrary rules to combine multiple different and imperfect tests inevitably creates a reference standard with some degree of bias. Importantly, all three studies in this group included the test under evaluation as part of the composite reference standard, which leads to additional bias, described below. Use of a biased composite standard is likely to lead to reduced sensitivity, among other errors affecting true test performance characteristics.

The fourth group of studies evaluated SARS-CoV-2 diagnostic tests that are under development as well as proprietary testing platforms (most of which are based on standard rRT-PCR methods). These studies used traditional rRT-PCR as a reference standard; results are summarized in Tables 4 and 5, respectively. Importantly, while these studies were not designed to estimate the accuracy of rRT-PCR, their results indicate that the index tests did not identify significantly more positive samples than rRT-PCR.

Finally, the last group of studies compared SARS-CoV-2 NAAT platforms (Table 6). These comparative accuracy studies examined the agreement between two non-reference standard tests. Although most of the testing platforms evaluated in these studies were based on standard rRT-PCR, the agreement between two non-reference standard tests is not equivalent to test accuracy, as mentioned previously.

This scoping review is limited by the lack of reporting of several key study features in the majority of the articles evaluated, which is an important indicator of quality and potential bias. Based on the QUADAS-2 criteria, most of the included studies had concern for bias (Supplementary Table 3). The most prominent concerns were unclear inclusion/exclusion criteria, unclear method of enrollment/selection of patients and samples, and unclear handling of indeterminate/inconclusive and invalid results. Additionally, many of the studies were conducted in a so-called “two gate” (case-control) design, in which cases and controls were known and selected ahead of time, rather than performing the test on a group of patients or samples with suspected COVID-19. These factors likely incorporate bias that significantly confounds the results of the studies, thus, the accuracy of the tests in other settings with different prevalences (such as asymptomatic screening, other age groups) may not be truly generalizable. Furthermore, few studies were able to evaluate both the index and reference tests simultaneously or within a short period of time, which is key to avoiding biases caused by changes in the patient’s true disease status; this bias can also affect the diagnostic accuracy of the index test.

The best approach to determining diagnostic test performance characteristics in the absence of a “gold” standard is an open question in diagnostic accuracy methodology. While many methods have been described, there are only a few well-defined statistical approaches that use a reference standard in lieu of a gold standard, reviewed elsewhere. Latent class analysis is one commonly used approach in situations in which neither the true error rates of the reference standard nor the true prevalence of the disease are known. This approach uses the results of a set of imperfect tests to estimate parameters related to sensitivity, specificity, and prevalence often using maximum likelihood methods. However, this is not the only method available and every method has its own strengths and limitations. Therefore, careful interpretation by studies that attempt to estimate test characteristics is warranted to account for and clarify the inherent limitations of assessing accuracy-related metrics when a gold standard is unavailable.

Evaluation of the performance characteristics of SARS-CoV-2 diagnostic tests is vital to control of the ongoing COVID-19 pandemic. While more than 200 SARS-CoV-2 molecular diagnostic tests have received FDA EUAs, we have described in this scoping review that the performance of few of these tests has been assessed appropriately. The lack of robust test performance that we noted in many studies published to date is undoubtedly due in part to the critical need for tests, which resulted in accelerated test development. However, our scoping review also uncovered imperfect methods for estimating diagnostic test performance in the absence of a gold standard and demonstrate that the accuracy of these tests should be interpreted with caution. Future studies would benefit from employing statistical methods such as latent class analysis and other methods referenced above to accurately analyze their data. Indeed, instituting national requirements for test performance analysis and reporting, perhaps based on the existing FDA guidelines on diagnostic tests, would advance the goal of standardizing the evaluation SARS-CoV-2 diagnostic test performance. Such an initiative would lead to statistically robust conclusions regarding the accuracy of the index test, which will in turn support hospitals and clinicians as they determine the optimal test to use for COVID-19 diagnosis.

Declaration of Competing Interest

None

Acknowledgments

None.

Funding

This work was supported by the National Institutes of Health [RO1CA232890] to E.Y.C.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jinf.2020.08.043.
References

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med* 2020; 382:727–33.
2. Lu R, Zhang X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and recombination. *Lancet* 2020; 395:565–74.
3. Wei W, Li Z, Chew J, Yong S, Toh MP, Lee VJ. Presymptomatic Transmission of SARS-CoV-2 – Singapore, January 23–March 16, 2020. MMWR Mortal Mortal Wkly Rep 2020; 69:411–15.
4. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med* 2009; 151:524–69.
5. Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, et al. Synthesis without meta-analysis (SWIM) in systematic reviews: reporting guideline. *BMJ* 2020; 369:s4690.
6. Whitaker PF, Burtis CA, Champion PW, Dongre M, Mallett S, Deeks JJ, Reitnsima JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011; 155:529–36.
7. Ai T, Yang Z, Hou H, Zhan C, Chen L, Li W, 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; 206:6042.
8. Liu R, Han L, Liu F, Li Z, Wu K, Liu Y, et al. Positive rate of RT-PCR detection of SARS-CoV-2 infection in 4880 cases from one hospital in Wuhan, China, from Jan 30 to Feb 4, 2020. *Clin Chem Acta* 2020; 505:172–7.
9. Xie C, Jiang L, Huang G, Hu P, Gong B, Lin H, et al. Comparison of different samples for 2019 novel coronavirus detection by nucleic acid amplification tests. *Int J Infect Dis* 2020; 92:57–64.
10. Bernstein A, Mei X, Huang M, Yang Y, Fayad ZA, Zhang N, et al. Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection. *Radiology* 2020; 295:20463.
11. Liu X, Zhang H, Xie J, Li M, Yang L, Pang P, et al. Sensitivity of Chest CT for COVID-19 compared with RT-PCR. *Radiology* 2020; 294:4032.
12. Green DA, Zucker J, Westblade LF, Whitther S, Rennert H, Velu P, et al. Clinical performance of SARS-CoV-2 molecular testing. *J Clin Microbiol* 2020; e00995–20.
13. He J, Liu L, Luo ZD, Lus JX, Ng MY, Shen XP, et al. Diagnostic performance between CT and initial real-time RT-PCR for clinically suspected 2019 coronavirus disease (COVID-19) patients outside Wuhan, China. *Respir Med* 2020; 168:109380.
14. Lee TH, Liu RJ, Liu RTP, Barkham T, Rao P, Leo YS, et al. Testing for SARS-CoV-2: can we stop at Two. *Clin Infect Dis* 2020.
15. Long C, Xu H, Shen Q, Zhang X, Fan B, Wang C, et al. Diagnosis of the Coronavirus disease (COVID-19): rt-PCR or CT? *Eur Respir J* 2020; 56:106061.
16. Wong HYF, Lam HYS, Fong AH, Leung ST, Chin TW, Lo CV, et al. Frequency and distribution of chest radiographic findings in COVID-19 positive patients. *Radiology* 2019; 201:1160.
17. Wu J, Liu J, Li S, Feng Z, Xiao Z, Wang X, et al. Detection and analysis of nucleic acid in various biological samples of COVID-19 patients. *Travel Med Inf Dis* 2020; 41:101673.
18. Cradick K, Lockhart M, Ozbolt P, Fatica L, Landon L, Lieber M, et al. Clinical evaluation and utilization of multiple molecular in vitro diagnostic assays for the detection of SARS-CoV-2. *Am J Pathol* 2020; 189:28–28.
19. Suo T, Liu X, Feng J, Guo M, Hu W, Guo D, et al. ddPCR: a more accurate tool for SARS-CoV-2 detection in low viral load specimens. *Emerg Microbes Infect* 2020; 9:1259–68.
20. Zhen W, Manji R, Smith E, Berry GJ. Comparison of four molecular in vitro diagnostic assays for the detection of SARS-CoV-2 in nasopharyngeal specimens. *J Clin Microbiol* 2020; 58:27.
21. Baek YH, Um J, Antigua KJC, Park JH, Kim Y, Oh S, et al. Development of a reverse transcription-loop-mediated isothermal amplification as a rapid early detection method for novel SARS-CoV-2. *Emerg Microbes Infect* 2020; 9:998–1007.
22. Kitagawa Y, Oirsha Y, Kawamura R, Imai K, Sakai J, Tanamoto N, et al. Evaluation of rapid diagnosis of novel coronavirus disease (COVID-19) using loop-mediated isothermal amplification. *J Clin Virol* 2020; 124:104446.
23. Lai YL, Ismail IM, Mustapa NL, Lai MY, Tuan Sth S, Hassan A, et al. Real-time reverse transcription loop-mediated isothermal amplification of rapid detection for SARS-CoV-2. *PeerJ* 2020; 8:e2978.
24. Lu R, Wu W, Wan Z, Li Y, Jin X, Zhang C. A Novel Reverse Transcription Loop-Mediated Isothermal Amplification Method for Rapid Detection of SARS-CoV-2. *Int J Mol Sci* 2020; 21:18.
25. Yan C, Cui J, Huang L, Du B, Chen L, Xue G, et al. Rapid and visual detection of 2019 novel coronavirus (SARS-CoV-2) by a reverse transcription-loop-mediated isothermal amplification assay. *Clin Microbiol Infect* 2020; 26:773–9.
26. Wang J, Cai K, He X, Shen X, Wang J, Liu J, et al. Multiple-centre clinical evaluation of an ultrafast single-tube assay for SARS-CoV-2 RNA. *Clin Microbiol Infect* 2020; 26:285.
27. Xue G, Li S, Zhang W, Du B, Cui J, Yan C, et al. A reverse-transcription recombinase-nested-aided amplification assay for rapid detection of the 2019 novel coronavirus (SARS-CoV-2). *Anal Chem* 2020; 92:22222.
28. Pereira R, Zhang C, Hao Z, Huang ML, Jerome KR, Greeninger AL. Multiplexing primer/probe sets for detection of SARS-CoV-2 by hRT-qPCR. *Clin Chem* 2020; 66:104499.
29. Waggoner JJ, Strittsburgh V, Pond R, Saklawi Y, Sahoo MK, Babiker A, et al. Detection of SARS-CoV-2 by real-time RT-PCR for severe acute respiratory syndrome coronavirus 2. *Emerg Infect Dis* 2020; 26:1633–5.
30. Li Y, Li J, Zhang Y, Dai L, Li L, Liu J, et al. Development of an automated inte-
mated system for rapid on-demand testing in the hospital setting. J Clin Virol 2020;128:104390.

56. Tsang TK, Wu P, Lin Y, Lau EHY, Leung CM, Cowling BJ. Effect of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China: a modelling study. Lancet Public Health 2020;5:e289–e296.

57. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess 2007;11 iii, ix-51.

58. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. Stat Med 1999;18:2987–3003.

59. Naaktgeboren CA, Bertens LC, van Smeden M, de Groot JA, Moons KG, Reitsma JB. Value of composite reference standards in diagnostic research. BMJ 2015;347:f5605.

60. Umemneku Chikere CM, Wilson K, Graziadio S, Vale L, Allen AJ. Diagnostic test evaluation methodology: a systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard - An update. PLoS ONE 2019;14:e0223832.

61. U.S. Food and Drug Administration CFDaRH. Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests. https://www.fda.gov/media/71147/download: [Accessed May 15, 2020]; 2007.