Guidance for the design and reporting of studies evaluating the clinical performance of tests for present or past SARS-CoV-2 infection

Doust, Jenny A; Bell, Katy JL ; Leeflang, Mariska M G; Dinnes, Jacqueline; Lord, Sally J ; Mallett, Sue; van de Wijgert, Janneke HHM ; Sandberg, Sverre ; Adeli, Khosrow ; Deeks, Jon; Bossuyt, Patrick M.; Horvath, Andrea R

DOI:
10.1136/bmj.n568

Guidance for the design and reporting of studies evaluating the clinical performance of tests for present or past SARS-CoV-2 infection

There is an urgent need for better guidance on the conduct and interpretation of diagnostic accuracy studies for SARS-CoV-2 tests so tests can be evaluated rigorously, but in an efficient and timely way. Even in a pandemic, clinical performance studies are essential prior to implementation.

Jenny A Doust[1], Katy JL Bell[2], Mariska MG Leeflang[3], Jacqueline Dinnes[4,5], Sally J Lord[6], Susan Mallett[7], Janneke HHM van de Wijgert[8], Sverre Sandberg[9,10], Khosrow Adeli[11,12], Jonathan J Deeks[4,5], Patrick M Bossuyt[3], Andrea R Horvath[2,13,14]

1. Centre for Longitudinal and Life Course Research, School of Public Health, The University of Queensland, Herston, 4006, Australia
2. School of Public Health, University of Sydney, NSW 2006, Australia
3. Department of Epidemiology and Data Science, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, Netherlands
4. Test Evaluation Research Group, Institute of Applied Health Research, University of Birmingham, UK
5. Public Health Epidemiology and Biostatistics, Institute of Applied Health Research, University of Birmingham, Birmingham, UK
6. School of Medicine, Sydney, The University of Notre Dame, Darlinghurst, NSW, 2010, Australia
7. Centre for Medical Imaging, University College, London, UK
8. Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands
9. Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway
10. Norwegian Quality Improvement of Laboratory Examinations, Haraldsplass Deaconess Hospital, Bergen, Norway
11. CALIPER Program, Paediatric Laboratory Medicine, The Hospital for Sick Children, Toronto, Canada khosrow.adeli@sickkids.ca.
12. Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Canada
13. New South Wales Health Pathology, Department of Chemical Pathology, Prince of Wales Hospital
14. School of Medical Sciences, University of New South Wales

Corresponding Author: Jenny Doust, Centre for Longitudinal and Life Course Research, School of Public Health, The University of Queensland, Herston, 4006, Australia; Email: j.doust@uq.edu.au. orcid.org/0000-0002-4024-9308

Jenny A Doust: Clinical Professorial Fellow

Katy JL Bell: Associate Professor in Clinical Epidemiology
Mariska MG Leeflang: Assistant Professor
Jacqueline Dinnes: Senior Researcher
Sally J Lord: Associate Professor
Susan Mallett: Professor in Diagnostic and Prognostic Studies
Janneke HHM van de Wijgert: Professor of Infectious and Immune-Mediated Disease Epidemiology
Sverre Sandberg: Professor and Director
Khosrow Adeli: Division Head and Professor
Jonathan J Deeks: Professor of Biostatistics
Patrick M Bossuyt: Professor of Clinical Epidemiology
Andrea R Horvath: Professor and Director

KB is supported by Australian NHMRC Investigator grant #1174523. SM is supported by grants from the National Institute for Health Research (NIHR), University College London (UCL) and UCL Hospital Biomedical Research Centre. SM, receives funding from the National Institute for Health Research (NIHR) and UCL/UCLH Biomedical Research Centre. KA is supported by grants from the Canadian Institutes of Health Research, Abbot Diagnostics, Siemens Healthineers, March of Dimes, and the Heart and Stroke Foundation of Canada. JJD and JD are supported by the NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. ARH has received a grant from the NSW Health COVID-19 Research Grants Round 1 – 2020: Improved Confirmatory Diagnosis of SARS-CoV-2 infection using Protein Mass Spectrometry

Funders had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Contributors: All authors provided a substantial contribution to the design and interpretation of the guidance, as well as writing sections of drafts, revising based on comments received, and approving the final version. ARH initiated the project. JAD wrote the initial draft, and is the guarantor for the study.

No patients were involved in setting the research question or outcomes of developing this study. No patients were asked to advise on interpretation or writing up of results.

Abstract

Testing for SARS-CoV-2 infection plays a key role in managing the current pandemic. This paper aims to provide guidance to assist researchers design robust diagnostic accuracy studies; publishers/peer reviewers to assess such studies; and to support clinicians and policy makers in their evaluation of the evidence on SARS-CoV-2 testing for clinical and public health decisions.

More than 1700 preprints and peer reviewed journal articles evaluating tests for SARS-CoV-2 infection have been published as of January 2021. However, evaluations of the studies to date have identified numerous methodological issues, leading to a high risk of bias and difficulties applying the results in practice. These problems demonstrate the urgent need for better guidance on the conduct and interpretation of these studies.

To address this need, this paper outlines the principles for defining the intended purpose of the test, the selection of study population, reference standard, test timing and other critical considerations for rigorous diagnostic accuracy study design, reporting and interpretation.

The implementation and accuracy of SARS-CoV-2 tests have major implications for individuals and communities, balancing the potential consequences of continued spread of infection against the need for public health measures, such as the restriction of movements and social activities. Making clinical and public health decisions in the current pandemic requires a clear understanding of the clinical performance and limitations of testing. We hope the current guidance will ensure that studies evaluating the diagnostic accuracy of SARS-CoV-2 tests are conducted using as rigorous methods as possible, in an efficient and timely way.

Introduction

Testing for infection plays a critical role in the response to the pandemic caused by the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) identified in China in December 2019[1]. Tests to identify SARS-CoV-2 infection and the disease caused by it (COVID-19) have been developed at an extraordinary pace; moving rapidly from the identification of the viral ribonucleic acid (RNA) sequence on 10 January 2020[2] to the development of viral nucleic acid tests for the virus using reverse transcription polymerase chain reaction (RT-PCR) shortly thereafter. This was followed by immunoassays for detecting the presence of viral antigens or antibodies in laboratories and at the point-of-care.

More than 1400 tests for SARS-CoV-2 are on the market or listed on websites such as the Foundation for Innovative New Diagnostics (FIND)[3] and the European Commission's Joint Research Centre database[4], and more than 1700 preprints and peer reviewed journal articles evaluating tests for SARS-CoV-2 infection have been published as of August 2020[5]. The volume of available diagnostic test accuracy evaluations is unprecedented and is unlikely to diminish with the implementation of programs to accelerate the development of new tests, such as the Rapid Acceleration of Diagnostics (RADx) program by the National Institutes of Health in the United States[6].

A vital part of managing the pandemic is to ensure that evaluations of tests for SARS-CoV-2 infection are rigorous, unbiased, and conducted in the most efficient way possible so that the most accurate tests are rapidly identified and adopted in practice. The evidence standards framework of the United Kingdom's National Institute for Health and Care Excellence (NICE) has outlined key evaluation concepts to assist with this process[7]. However, systematic reviews of diagnostic accuracy studies of tests for SARS-CoV-2 have highlighted many methodological and reporting problems (Table 1)[8-14]. These problems limit the ability of clinicians and policy makers to apply the results of such studies in diagnostic pathways and public health programs and have led to poor clinical and public health decisions contributing to ongoing spread of the infection[15].

The objective of this paper is to outline general principles for studies that evaluate the clinical performance of SARS-CoV-2 tests. For ease of reading, we use the term "SARS-CoV-2 tests" when we are referring to any of the following: viral nucleic acid, antigen or antibody detection tests. The authors have expertise in the evaluation of diagnostic tests including the evaluation of SARS-CoV-2 tests, evidence-based medicine, epidemiology, laboratory medicine, and virology. We have based the guidance in this paper on previously published work on diagnostic test evaluations, such as the STARD guideline for reporting of diagnostic accuracy studies[16], and the QUADAS-2 tool for appraising the risk of bias of diagnostic accuracy studies[17]. We have also considered the guidance provided in templates issued by the Food and Drug Administration (FDA) for Emergency Use Authorizations for *in-vitro* diagnostic tests for SARS-CoV-2[18], the NICE evidence standards framework[7], the Medicine and Healthcare products Regulatory Agency (MHRA) and World Health Organization target product profiles[19, 20], and the European Commission's document on COVID-19: Recommendations for testing strategies 2020[21] and related documents[22].

The focus in this paper is on clinical performance studies addressing the diagnostic accuracy of SARS-CoV-2 tests in clinical or public health practice. Many of the studies initially undertaken and quoted in reports of test performance can be classified as studies of scientific validity (see Box 1).[23] They are essential in the development of a test, analogous to the finding of Phase 1 clinical trials. Similarly, analytical performance studies, are also necessary prerequisites before clinical application of a test[22]. These studies cannot, however, provide realistic estimates of the diagnostic accuracy of the tests when used in clinical practice and it is misleading to assume the results from such studies apply in the clinical setting.

Our test evaluation guidance is outlined in a series of steps, in the order of the STARD checklist, although the steps may not be sequential in practice. Table 1 outlines the STARD checklist items, noting some key methodological issues in the studies done of SARS-CoV-2 tests to date. The steps described below are illustrated with examples of possible study designs in Table 2.

Step 1: Define the intended use of the test

Many published evaluations of SARS-CoV-2 tests are not able to provide an accurate estimate of the performance of the test in clinical practice because the relationship between the purpose of the test, the selection of the study population, and the selection of the reference standard have not been carefully mapped out prior to the conduct of the study. Before beginning an evaluation of a SARS-CoV-2 test, it is important to define how the test will be used in the clinical or public health pathway. Some possible indications for use of SARS-CoV-2 tests are[24].

For viral nucleic acid (such as RT-PCR) and antigen testing:

1) to diagnose COVID-19 in individuals with symptoms suggestive of the disease;
2) to test asymptomatic, pre-symptomatic or mildly symptomatic individuals with known recent exposure to a confirmed case (e.g. as part of localised outbreak investigations and test-and-trace programs);
3) to screen individuals at risk of acquisition and/or transmission of infection (e.g. staff/patients in hospital or staff/residents in aged care or education facilities, as part of outbreak prevention programs);
4) to evaluate if a person with SARS-CoV-2 infection has cleared the virus;
5) to establish the prevalence of current SARS-CoV-2 infection in a population (e.g. for public health decisions, or to estimate pre-test probability for an individual in that population);

For serology (antibody) testing:

1) to investigate patients presenting late after symptom onset in whom viral nucleic acid testing is negative or where viral nucleic acid testing is not available to confirm whether they were infected with SARS-CoV-2;
2) to determine antibody presence as part of a broader immunological assessment (e.g. in intervention studies evaluating the efficacy of SARS-CoV-2 vaccine immunogenicity or convalescent plasma).
3) to estimate the seroprevalence of past and recent SARS-CoV-2 infection in a population (e.g. for public health decisions);

Testing to assess if an individual has immunity to further infection is also of key interest. However, this requires studies that demonstrate that specific immune responses, such as the presence of antibodies (neutralising or non-neutralising), T cell and/or other cellular responses, lead to protection from clinically important infection or re-infection. The detection of antibodies *per se* is insufficient to demonstrate immunity. As yet, we do not have strong evidence of what immune responses are necessary for immunity to SARS-CoV-2 infection[25-27].

Defining the clinical (or public health) pathway involves not only describing the test, but also the test population, the role and position of the test (including what tests are conducted before and after the test being studied), how the test results will be used and their impact on management decisions.

Testing strategies also need to consider the availability of test materials and other resources, and the prevalence of infection in the community. Each type of test has different requirements in terms of equipment, expertise of the operator, sample types, sample storage, and turn-around time. Mathematical modelling studies have shown that reducing the time between symptom onset and a positive test result, assuming immediate isolation, is the most important factor for improving the effectiveness of test-and-trace programs[28], so in some settings there may be a trade-off between turn-around time and diagnostic accuracy.

False negative test results may lead to infected individuals continuing to come into contact with and potentially infecting other individuals. False positive test results may lead to individuals being told incorrectly that they are infected with SARS-CoV-2 and decisions regarding isolation measures, restriction of movement and activities for both the individual and the community. The rate of infection in the group (i.e the prevalence in the group) will impact on the predictive values of the test, that is the probability of false positive and negative test results (Figure 1). For example, in settings where there is a very high rate of transmission, the pre-test probability of infection for an individual may be so high that even a negative test result does not safely rule out infection to a level that an individual can be assumed to be non-infectious unless the test has a very high sensitivity[29].

Groups such as the FDA in the United States[18] and the MHRA in the United Kingdom[19] and the World Health Organization[20] have set acceptable and desirable performance characteristics for SARS-CoV-2 testing (called target product profiles in the case of the latter two). The targets set by these agencies show a low tolerance for both false negative and false positive results in the setting of the SARS-CoV-2 pandemic. Acceptable clinical performance characteristics are determined by the values placed on the consequences of testing and are not definitive or intrinsic to the test.

Where clinical pathways are more established, it is generally desirable to establish minimum acceptable clinical performance characteristics prior to conducting a clinical performance study[30]. In the setting of a pandemic, however, where the rate of infection in the community is changing and new tests, treatments and responses to infection are rapidly becoming available, this is not likely to be feasible. In this context, groups conducting clinical performance studies should make the information from their protocols and reports available to public health and clinical decision makers in a rigorous, transparent and timely manner.

Studies should also clearly outline existing or alternative clinical pathways, including whether the test being evaluated is intended to replace an existing test or is in addition to existing testing[31]. For example, a reverse transcription loop-mediated isothermal amplification (RT-LAMP) test might be used as a replacement diagnostic test for RT-PCR, to reduce the demand for reagents and allow for faster turn-around-time. Studies that explicitly compare diagnostic tests in clinical pathways are valuable for clinical and public health decision makers.

Understanding the timing of the viral and immunological responses to SARS-CoV-2 infection is critical in considering the clinical pathway. After exposure to SARS-CoV-2, the virus typically becomes detectable by RT-PCR testing on the 3rd or 4th day after infection (Figure 2)[32, 33]. Symptoms typically appear around the 5th day of infection, and both symptoms and viral detection last for several days to weeks, depending on the severity of infection[34]. Studies using repeat RT-PCR testing and tracking of transmission rates (including infector-infectee transmission pairs) have shown approximately 40% of transmissions occur prior to the development of symptoms[35], and peak infectivity occurs approximately one day prior to until 2-3 days after symptom onset in typical individuals[32]. Antibodies are generally low in the first week after symptom onset (in cases confirmed with RT-PCR), with most individuals seroconverting by day 10 to 14, and diagnostic sensitivity for SARS-CoV-2 infection of serology tests only exceeds 90% in the third week after symptom onset[8-10], and then begins to decline[36]. It is not yet known how long high levels of antibodies to SARS-CoV-2 infection persist, but the observations to date show that the response among individuals varies, influenced by disease severity[26,27,36].

It may not be possible for researchers to predict all aspects of intended uses of the test as well as consequences of the test result. However, it is important that researchers consider the potential clinical pathways *a priori* and how this will affect the application, timing and interpretation of the results of the test, and therefore the design of their study.

Step 2: Define the target condition

Building on the first step, it is important to clearly define the target condition of interest; that is what the test aims to detect. For SARS-CoV-2 tests, potential target conditions include infection with the virus, disease caused by the virus (that is COVID-19), the extent of infectivity, the presence or extent of immune responses to the virus, clearance of the virus, past or recent infection with the virus, and immunity to infection. Explicit consideration of the target condition(s) of interest helps identify further elements that guide study design, such as the population to be tested and acceptable reference standards for defining the presence of the target condition. For most clinical performance studies, the target condition will be SARS-CoV-2 infection (which includes symptomatic, pre-symptomatic and asymptomatic infection).

In some settings, it may be more important to establish if someone is infectious rather than if someone has the infection. For example, if an individual presents in a healthcare setting, knowing if they are infectious or not influences the need for personal protective equipment and other infection control measures immediately, whereas determining if they have the infection is less urgent if the individual’s symptoms are mild but SARS-CoV-2 infection cannot be excluded. Testing for infectivity, rather than infection, has also been suggested as a possible method for screening in other settings, including opening businesses and allowing public gatherings[37]. While such strategies should be investigated, the entire clinical pathway for such strategies needs to be evaluated, including the potential consequences of false positive and negative test results.

Step 3: Define the population in which the test will be evaluated

Poor patient selection and description of study groups have severely limited the ability to establish the diagnostic accuracy of SARS-CoV-2 tests to date. Scientific validity studies, often of a case-control design, cannot provide realistic estimates of the diagnostic accuracy of the tests when used in clinical practice. To establish diagnostic accuracy, clinical performance studies should be conducted in individuals sampled from the population in which the test will be used, as determined by the intended use in Step 1. Examples of possible populations for diagnosing current (or prior) infection include: individuals with current (or previous) symptoms suggestive of COVID-19; individuals at high risk of exposure (such as close contacts of confirmed cases); individuals at high risk of both exposure and transmission (such as healthcare workers or residents of aged care facilities) and patients hospitalised with suspected COVID-19. Based on the target population, studies should then define the method for enrolling participants into the study, including inclusion and exclusion criteria, aiming to recruit participants representative of the target population. Ideally, where the intended test use is in a healthcare setting, consecutive individuals from the target population would be recruited without prior knowledge of whether the individuals have the target condition or not. For population-based studies, where the intended test use is for public health decisions, a representative random sample of the target population could be used. Studies using known cases and healthy controls introduce selection bias and effects related to the clinical spectrum of disease.

The diagnostic accuracy observed in studies conducted in patients hospitalised with severe COVID-19 or recruited from hospital settings may not be applicable to other settings. For example, although the intended use population for most serology tests is a community setting that includes individuals who have experienced no or mild COVID-19 symptoms, most published studies of these tests have recruited patients hospitalised with severe infection. Antibody production in this population is likely to be higher than in the wider population of those infected[8].

If the purpose of the test is to establish the presence of SARS-CoV-2 infection in a community setting or a clinical population, patients with respiratory symptoms due to respiratory illnesses other than SARS-CoV-2 should not be excluded from the study as these patients will be tested in clinical practice. Careful thought needs to be given to the presence or absence of symptoms that may be used as eligibility criteria for the study. The presence of, for example, respiratory symptoms prompts correct selection of the anatomical site for the sample and correct timing (during symptoms). When testing for asymptomatic infection, neither of these helpful prompts are available, meaning that other epidemiological information (e.g. risk of exposure, and time since exposure, if known) and more than one sample (anatomical or time point) may need to be tested. Viral nucleic acid is typically able to be detected on the 3rd day after exposure in nasal, throat or saliva secretions[32,33]. It is unclear if virus is typically detected in faeces and sputum at 2 days post infection or if later time points are relevant for these sites of sampling.

In addition to defining the population, when conducting the study it is important to record and report characteristics of study participants during the course of the study, such as the presence of key symptoms (temperature, cough and so forth), time since a high-risk contact (defined as contact within a certain distance of a person with confirmed or probable SARS-CoV-2 infection and for a certain amount of time), viral load if known, markers of disease severity and time since the development and cessation of symptoms. The number and reasons for any exclusion of individuals from the study following recruitment should also be recorded.

As the accuracy of all tests depends on their timing, it is essential to record the time point in the disease course at which the test is done, in relation to time since known exposure and time since onset of symptoms. Due to differences in health care provision and pathways, only recording time since health care events (such as admission to hospital, ICU, or results from RT-PCR) restricts the ability of study findings to be generalised to other settings.

Step 4: Describe the index test

Given the natural history of infection over time, variations in viral load, and the current limitations in test accuracy, combinations of tests and/or tests at different time points may be needed to identify all true cases and non-cases. The index test strategy may therefore be a single test, or the same test repeated at different time points or a combination of different tests, such as a test with lower specificity followed by a test with higher specificity in those initially positive. Ideally, the entire testing pathway would be evaluated.

SARS-CoV-2 tests may be developed commercially or “in house” by a laboratory and need to meet key regulatory and/or emergency use authorization requirements for *in vitro* medical devices[18-22]. All pre-analytical, analytical and post-analytical characteristics of the test should be described, including

- the full name of the test and manufacturer, and associated batch numbers allowing clear identification;
- pre-analytical characteristics
    - type of samples suitable for testing (e.g. nasopharyngeal swab, sputum, saliva, blood),
    - method of collection of specimens and how the sample was taken (e.g. whether a long swab was used for RT-PCR tests),
    - who has taken the sample (such as their clinical training),
    - conditions for specimen handling, transport and storage;
- analytical characteristics
    - actual target of the assay (what is being measured, e.g. viral nucleic acid, or antigen, or antibody against specific viral proteins),
    - principles of analytical methods (e.g. fluorescent, multiplex fluorescence or digital RT-PCR; enzyme-linked immunoassay or lateral flow assay),
    - the platform used for measurement (how and with what device the target analyte is measured),
    - where was the analysis done, if relevant (for example at the point-of-care or in a reference laboratory),
    - the analytical performance measures of the test (e.g. analytical sensitivity/limit of detection, cross-reactivity, accuracy, trueness, precision);
- post-analytical characteristics
    - test interpretation,
    - decision limits at which the test is considered positive or negative, where applicable.

Pre-analytical Characteristics - Specimens
The study should determine *a priori* which specimen types will be tested. The results of evaluations on one type of specimen cannot be generalised to other specimen types without further validation. The type of specimen and the methods used to collect and analyse the specimen need to reflect the methods intended to be used in standard clinical practice. For PCR and antigen tests, the anatomical site used for collection of the specimen should be stated; for example, whether the specimen is taken from the upper respiratory tract (nasal or pharyngeal swab – including insertion depth, or saliva), from the lower respiratory tract (bronchoalveolar lavage, sputum) or other (urine, faeces, blood). Samples using viral transport medium spiked with inactivated virus are not appropriate for assessing the test's clinical performance. For antibody tests, the sample type could be venous whole blood, plasma, serum, or finger-prick capillary whole blood. Elution protocols for dried blood spots should be available if used. Tests should be evaluated preferably with samples that are prospectively collected.

Analytical Characteristics
The actual targets that the test is measuring must be clearly stated or reference must be given to the actual measurement procedure or vendor's instructions. For viral nucleic acid tests by RT-PCR, the primer binding site/s, and for antigen tests, the specific antigen targeted should be stated and whether the specimens were run with or without extraction, heat inactivation or pooling. For serology tests, it is important to describe the viral proteins targeted by the antibody (typically the Spike protein, S1 or S2, which are specific for SARS-CoV-2, and/or the nucleocapsid protein, which is conserved among all coronaviruses); the type of immunoglobulin(s) detected (i.e. IgA, IgG, and/or IgM); and the immunological method used (e.g. enzyme-linked immunosorbent assay [ELISA], chemiluminescence immunoassays [CLIA], lateral flow immunoassays [LFIAs], and fluorescent immunoassays [FIA]). Depending on the question being asked as determined in Step 1, the authors will also

need to determine whether the index test is identifying neutralising or non-neutralising antibodies.

The key analytical performance indicators of the tests used in the evaluation should be known before starting a clinical performance study. These characteristics should be described, if possible, using appropriate reference measurement methods to ensure they adequately measure the presence and/or quantities of the virus or antibodies, and will usually be described in the instructions for use documentation. These typically cover the limit of detection, reportable range, imprecision, trueness as compared to a reference method and the analytical specificity of the tests. Recommended methods for performing these analyses are given in the FDA templates[18] and elsewhere[38,39]. Quality controls, such as negative and positive controls and linearity checking by measuring of levels using spiked samples with increasing concentrations of the virus, antigen or antibody, are also necessary. For RT-PCR, the limit of detection is typically measured by spiking RNA or inactivated virus into an artificial or real clinical matrix, such as bronchoalveolar lavage fluid or sputum. The limit of detection should be reported, for example as viral copies/mL.

Cross-reactivity with other viral RNA or antigens or antibodies to previous infections (analytical specificity) also needs to be evaluated to show that the test does not cross-react with normal microbiota or other pathogens that may be present in the clinical specimen. High priority organisms for the evaluation of cross-reactivity are listed in the FDA templates[18]. Potential cross-contamination within the laboratory also needs to be minimised, but this needs to be controlled by good laboratory practice to avoid carry-over. Contaminated reagents in laboratories have led to false positive test results[40]. A proportion of samples within the study should therefore be tested for cross-contamination/carry-over and this proportion should be stated.

Measures of precision (repeatability and reproducibility) may be important, for example if different operators will be analysing results in the laboratory or at the point-of-care. Repeatability reflects closeness of agreement between results of successive measurements carried out under the same laboratory conditions, while reproducibility reflects closeness of agreement between results of measurements performed under changed laboratory conditions of measurements (e.g., time, operators, calibrators, and reagent lots)[41]. The lot-to-lot variability of tests, such as lateral flow assays, should be stated.

Postanalytical Characteristics - Decision Limits

Decision limits need to be defined for positive, negative and indeterminate results. Preferably, these cut-points are selected *a priori*, for example based on the manufacturer's guidance, or from previous scientific validity studies. If invalid or indeterminate results are repeated, the methods for deciding this process should be described and the number of such repeat tests should be reported. Cut-points derived from the data collected within the study can bias estimates of test performance[42,43]. If no prior data exist to determine cut-points, or when the cut-point was established in symptomatic cases but the test is intended to be used in non-symptomatic or mildly symptomatic individuals, then it must be made clear that further external validation of the optimal cut-point is needed in an appropriately selected and representative population.

For RT-PCR tests, there has been considerable discussion regarding the number of amplification cycles used and the cycle threshold (C_T) to determine if a test is positive, negative or indeterminate. While there is a strong relationship between C_T and viral load, choosing the C_T is not easily generalizable between tests, kits, testing platforms and laboratories. C_T values may be transformed into concentrations using a calibration curve for each testing pathway (test, kit, platform and laboratory), allowing for direct comparisons between different testing pathways. The C_T or concentration cut-offs used in the evaluation should be clearly explained, and the methods for managing an indeterminate test clearly outlined.

Step 5: If applicable, describe which tests are compared and why

With the rapid development of so many SARS-CoV-2 tests, decisions need to be made regarding the comparative performance of different tests. The comparison may be between different forms of testing or different tests of the same form or different testing strategies. Each test included in the study should be described as in Step 4.

Comparisons of index tests may involve a comparison of two or more index tests against a common reference standard or compare the agreement of two tests against each other. In the case of the former, it is preferable that both index tests are performed in the same individuals, using a direct comparison, rather than an indirect comparison of the index test against the reference standard in two different study groups.

Studies that make head-to-head comparisons of many tests in the same samples efficiently provide important and useful information about comparative test accuracy. However, the practicalities of obtaining adequate samples to perform all included tests without compromising the generalisability of the study findings must also be considered.

The aim of the comparison should be specified. For example, the aim of the study may be to perform a descriptive analysis of all included index tests or may be to determine if a new test has higher sensitivity and equivalent specificity, or faster turn-around time and equivalent diagnostic accuracy. Although one characteristic may be specified as the primary outcome, for example improved sensitivity, other measures of clinical performance will also need to be evaluated, such as the test's specificity. Note that the comparator test is not the same as the reference standard described in Step 6.

Step 6: Define the reference standard

The reference standard needs to clearly separate those who have the target condition from those who do not; for example, those who have/have had the infection from those who do not/have not had the infection, or those who are infectious from those who are not infectious. Irrespective of the intended use, in clinical performance studies, the interpretation of the index test/s, the comparator test/s, and the reference standard test need to be conducted masked to the results of the other test/s.

In the systematic reviews of SARS-CoV-2 tests to date, a high proportion of studies have used a reference standard with a high risk of bias, and that is not applicable to the clinical population of interest[8-14]. Selection of the appropriate reference standard for evaluation of SARS-CoV-2 tests is not simple, and several issues described below need to be considered[44].

For studies where the target condition is SARS-CoV-2 infection

Cases of SARS-CoV-2 infection include individuals who are asymptomatic, pre-symptomatic and symptomatic. The World Health Organization (WHO) has published definitions of suspected, probable and confirmed COVID-19 cases based on clinical, epidemiological and laboratory criteria, with recommended associated testing[45,46]. According to this advice, a confirmed case of COVID-19 is defined as a person with laboratory confirmation of SARS-CoV-2 infection, irrespective of clinical signs and symptoms. This unfortunately generates some confusion as in most publications COVID-19 is the disease caused by the SARS-CoV-2 virus and thus is equivalent to symptomatic infection, not to infection *per se*.

The WHO defines a probable case of COVID-19 as an individual who has symptoms indicative of the disease (fever, cough, general weakness/fatigue, headache, myalgia, sore throat, coryza, dyspnoea, anorexia/nausea/vomiting, diarrhoea, altered mental status), AND has an epidemiological risk of exposure AND: a) is a contact of a probable or confirmed case; or b) has chest imaging findings suggestive of COVID-19; or c) has a loss of taste or smell; or d) death has occurred that is not otherwise explained in an adult with respiratory distress preceding death and was a contact of a probable or confirmed case or epidemiologically linked to a cluster with at least one confirmed case. The WHO definitions above are necessary to standardise clinical protocols and reporting but will also misclassify a proportion of cases. Some individuals will be classified as a case, mostly as a probable case, who are not infected with SARS-CoV-2. On the other hand, some individuals have had exposure, have had symptoms and investigations such as imaging that indicate COVID-19, but testing (either RT-PCR and/or antibody) has been negative. These individuals are not classified as definite cases. If the WHO classification is used as a reference standard, it is helpful to present a sensitivity analysis of clinical performance of the test using a reference standard including probable cases of disease.

Putting aside the confusion caused by terminology, viral nucleic acid testing (specifically RT-PCR) is frequently used as a reference standard for SARS-CoV-2 infection, where the individual has had possible exposure up to 2 weeks prior to testing. After this period, viral load decreases in many individuals reducing the sensitivity of the RT-PCR. Although it is thought that the specificity of viral nucleic acid testing is very high, it is not 100%. The probability of false positive test results is difficult to determine, but it is possible that at least some individuals who have tested positive and who remain asymptomatic have never had the virus. Some false positive test results may be due to cross-contamination with other samples or clerical error in reporting results. Repeat testing may identify some false positive results, but interpretation of discordant results is complex. For example, a second test, especially if done beyond the typical 14 days test window post-exposure, may be negative because the individual is no longer viraemic. Repeat testing in individuals with confirmed COVID-19 shows that false negative results also occur, particularly in the first few days after exposure or late in the course of infection[32-34,47,48]. Poor sampling technique, samples from the “wrong” anatomical site and incorrect transport of specimens can also contribute to false negative results. A single negative viral nucleic acid test is inadequate to rule out SARS-CoV-2 infection.

Performance of viral nucleic acid testing as a reference standard may be improved by ensuring appropriate collection, repeat testing for those who initially test negative within an appropriate time window (for example, within 5 days post symptom onset or on the fourth day post exposure if exposure date is known), or by samples from multiple sites or with multiple genetic targets[49,50]. Serology may be used where it is thought that exposure may have occurred more than 14 days prior. However, it also has a high false negative rate, and may also have false positive results due to the presence in the specimen of substances such as rheumatoid factor, heterophile antibodies, haemolysis, fibrin, and other types of coronaviruses,[51,52] or from an earlier SARS-CoV-2 infection. Repeat testing and combinations of tests, however, involves a greater layer of complexity in deciding what is considered a true positive and true negative result and will add to the resources needed to conduct an evaluation. If repeat or multiple testing is used as part of the reference standard, the testing strategy needs to be clearly outlined with the same strategy used for all individuals included in the study, not just those samples where there is a discordant result between the index test and the reference standard[53].

For asymptomatic infection, clinical reference standards are not possible as there are no clinical symptoms and the number of asymptomatic patients detected with other forms of testing, such as lung imaging to detect inflammation, will be low.

For studies where the target condition is COVID-19
COVID-19 is the disease caused by SARS-CoV-2 and therefore includes all patients with symptoms. For diagnosing COVID-19 disease, the clinical reference standard is likely to be a combination of clinical information, including repeat/multiple RT-PCR tests, other tests (including chest imaging), serological antibody testing and clinical follow-up. Studies should specify which clinical information is used as part of the clinical reference standard and attempts made to obtain this information for all study participants, for example using the information included in the WHO definitions for probable cases. Clinical follow-up and repeat testing of those who develop symptomatic disease or more severe disease will detect at least a proportion of individuals with COVID-19 who are initially negative on RT-PCR testing[12]. The use of multiple sources of clinical information as a reference standard ensures more complete identification of cases, but it can also lead to both an under-estimation of the diagnostic sensitivity of an index test (if individuals are defined by the reference standard as cases of disease are actually true negatives) or an over-estimation of the sensitivity of an index test (if the results of the index test are incorporated into the definition of the target condition). A reference standard using all clinical information, while not perfect, is probably the best that can be achieved at present.

For studies where the target condition is previous SARS-CoV-2 infection
If the purpose of the test is to identify previous SARS-CoV-2 infection, for example to validate use of a serology test for a seroprevalence survey, the reference standard needs to demonstrate clear evidence of the presence or absence of previous infection. This may be done through results of a prior RT-PCR test plus clinical information about potential exposure risk and clinical follow up. Timing of such testing with RT-PCR is difficult, especially in asymptomatic and pre-symptomatic cases. Therefore, if the test is intended for seroprevalence surveys, the best study design would involve a large number of randomly selected cases who are regularly tested with repeat PCR weekly or biweekly as a reference standard and followed up by serology testing 2-3 weeks after the last RT-PCR test until there is risk of exposure to the virus. However, such studies, especially in a low prevalence setting, would be costly and uncomfortable to study participants.

Exclusion of prior infection needs to be established as robustly as the presence of current infection. Many studies evaluating serology tests have used samples from pre-pandemic serum and blood banks, either from health resources or from study sample archives. Such studies can measure scientific validity and analytical sensitivity and specificity, but do not measure clinical performance.

Comparisons of different forms of serology testing can be valuable, but must be made against an appropriate reference standard, and require understanding that the development of an immune response varies between individuals in the timing, intensity and which parts of the virus antibody responses are targeted. Inclusion of a probable case category may be of use.

For studies where the target condition is infectivity
Although a positive RT-PCR test result indicates presence of viral RNA, it does not necessarily indicate that the individual is infectious. Infectiousness requires the virus to be present in a bodily secretion that could result in transfer of virus to another individual, and also that the virus particles in secretions remain infectious i.e. are still viable virus particles as opposed to inactive or remnants of virus particles. The ability to use a rapid test that determines if an individual is infectious could have advantages in some settings, as described above. However, a reference standard for determining viable and non-viable viruses in the patient's specimen does not currently exist. Assays of virus infectivity in cell culture and viral replication could be a measure of virus viability and infectivity, but are currently not suitable outside a research setting, as the assays are time consuming and methods are still being refined including sampling methods, transportation and culture

media. Cell culture assays are problematic as a reference standard as they appear to have suboptimal sensitivity for detecting infectivity. Early in the course of infection, when we expect most cases are infectious, samples from RT-PCR positive cases may not grow virus on cell culture[54]. While samples that return a positive result at a higher C_T may indicate viral remnants at a point where the patient is no longer infectious, they may also indicate an early point in the course of the infection, and reducing the C_T will reduce the sensitivity of the test to detect infectious individuals. Similarly, assuming that only those with high viral load are infectious will miss individuals who have lower viral loads but are still capable of passing on the infection[15].

For studies where the target condition is SARS-CoV-2 infection clearance
Diagnosis of SARS-CoV-2 clearance (i.e. absence of detectable viral particles whether viable virus or not) generally requires at least two negative RT-PCR tests to demonstrate clearance. However, testing at multiple anatomical sites has shown that the virus is cleared from the upper respiratory tract before clearance from the lower respiratory tract[11]. Time for clearance from gastrointestinal tract varies greatly by individual. It is unclear if the presence of the virus in faeces has a role in the spread of infection, although this was a significant route for spreading infection in SARS.

Step 7: Analysis and presentation of results
Poor reporting of studies evaluating SARS-CoV-2 tests has been a common methodological concern in the studies to date. Reports should follow the STARD reporting guidelines for diagnostic accuracy studies[16]. This includes the STARD flow diagram to report the number of individuals included in the study, the number of individuals excluded from the study prior to testing, the number of individuals whose samples were not tested and the number of individuals who had samples tested but who were not included in the study (for example, who did not receive the reference standard, or had indeterminate or outlier results) (Figure 3). The diagram may need to be adapted in the case of studies that use repeated testing over time. The prevalence of SARS-CoV-2 in the study group needs to be clearly identified, and where possible, study reports should indicate transmission intensity and co-circulating pathogens at the time of the study.

Sample size and unit of analysis
The sample size should be the number of individuals included in the study, not the number of samples tested. If more than one test from some individuals are included in the study, the repeat test should not be included in the same estimates of sensitivity and specificity. Repeat samples from the same individual can be included, however, for the estimation of sensitivity and specificity at different time points (one repeat at each time point). Such analyses can be helpful in establishing the sensitivity and specificity of a test over time. Where repeat testing occurs, the reason for repeat testing should be reported and the reporting of repeated samples should be clear. If more than one test from all individuals are included in an evaluation of a testing strategy (rather than evaluation of a single test), then the sample size is again the number of individuals included in the study).

Although it is important to conduct evaluations of sensitivity and specificity in the same population to estimate clinical test performance, preliminary studies may estimate sensitivity and specificity in separate study groups. Where this occurs, the sample size for each group should be stated separately.

Analysis of data
In presenting the results of the study, it is helpful to provide a cross-tabulation of the index test and the reference standard results. Using the same reference standard for all index tests minimises the risk of verification bias. Where there are missing data or indeterminate

results for either the index test or reference standard, these should be reported according to the final disease status (if known) and not excluded from the results.

Reports may include the results of analytical performance (analytical sensitivity, analytical specificity, imprecision, etc), but these need to be clearly differentiated from clinical performance (diagnostic or clinical sensitivity and specificity) which are the more relevant measures and should be the focus of the report. All estimates require confidence intervals, based on the appropriate sample size using appropriate methods for computation, such as exact binomial or Wilson approximation[55,56].

Timing
For each individual included in the study, the timing of the samplings and the analysis of the test should be recorded. Time from presumed exposure to infection and since the onset of symptoms (if applicable) should also be recorded. In general, the index test and the reference standard should be conducted as close in time as possible. If both the index test and the reference standard include RT-PCR, then the same sample should be used or paired samples should be obtained.

For studies evaluating antibody tests to identify previous infection, the reference standard may include a RT-PCR test and/or other tests conducted during the symptomatic phase of the illness or post-exposure, with antibody testing conducted at a later date, when the individual is likely to have seroconverted. In these studies, the timing of the serology sampling may be defined as time since RT-PCR evaluation, or better, the time since exposure to a known case or since onset of symptoms. For studies using a reference standard that includes clinical follow up or repeat testing, the same follow-up period should be used in all individuals included in the study.

Sub-group analyses
Sub-group analyses of diagnostic performance by factors known to affect the sensitivity and specificity of testing can assist the understanding of the clinical applicability of the results. Most of the identified heterogeneity for SARS-CoV-2 tests seen so far is in the sensitivity of the test. Sub-group analyses by time since exposure, time since symptom onset, disease severity, viral load or antibody titre in the reference standard and in groups of individuals who are asymptomatic/pre-symptomatic or symptomatic are particularly helpful.

Comparative analyses
As described above, ideally two index tests will be compared within the same study group. Where two index tests are measuring a common property and no reference standard is used, the agreement between tests may be reported in the form of tables showing concordant and discordant results. Further information on the people with “discordant” results may help to evaluate which test is more accurate using agreement with observations that may be considered “fair umpires” but are not a reference standard[57]. Such “fair umpires” could include information on prior exposure risk, concurrent tests (besides index or comparator test under evaluation, e.g. inflammatory markers, chest imaging), response to treatment, and clinical outcomes on follow-up.

Predictive values
Clinicians and public health experts require an understanding of the positive and negative predictive values of the test, not just the sensitivity and specificity of the test. In presenting the results of the study, it may be helpful to provide estimates of these using several clinically relevant values of prevalence. It is also helpful to display how the test characteristics will perform in different prevalence settings graphically, and using natural frequencies (such as the number of people affected in a population of 10,000 people), as shown in Figure 1. A calculator to convert sensitivity, specificity and prevalence to the

positive and negative predictive values of the test that are relevant to the target population is provided on the FDA website[58].

In addition to summarising the results, authors can provide guidance to assist those using the study results (such as clinicians, public health staff and policy makers) on how the results of the study can be applied in practice and the consequences of false positive and false negative test results. Where possible, advice can be given on how testing strategies and use of the test may need to be refined based on the understanding gained from the evaluation of the test.

If a study is done in a reference laboratory with highly experienced staff, it needs to be acknowledged that the results will represent the best-case scenario for the estimates of diagnostic accuracy, and the test is likely to have performance characteristics that are less than this in clinical practice.

If future research is needed, advice on how to store samples and how to assure the stability of samples and what data to record for biobanking purposes can be helpful. Appropriately designed and harmonised sample banks, with detailed information about the population characteristics, should be made available to developers of new tests so that the tests can be rapidly validated, and passed to clinical laboratories for local verification.

Step 8: Prospectively register the study protocol

On completion of the study design, study protocols can be registered before their initiation in a clinical trial registry, such as ClinicalTrials.gov or one of the WHO Primary Registries ensuring that existence of the studies can be identified[59]. Prospective registration is a sign of quality, providing evidence that the study objectives, test procedures, outcome measures, eligibility criteria and data to be collected were defined prospectively, and allows transparent reporting of any modifications to study protocols. Trial registration also allows reviewers to identify studies that have been completed but were not yet reported, supporting the reduction in publication bias in subsequent systematic reviews. Including a registration number in the study report facilitates identification of the trial in the corresponding registry.

Conclusion

Testing and early identification of individuals with SARS-CoV-2 infection is a vital part of controlling the spread of the pandemic, including decisions regarding the need to introduce public health measures such as restrictions on movements and limits on social gatherings. To do this, we need to establish the clinical accuracy of tests in rigorously designed evaluations and in the full range of intended use settings so that the consequences of acting on test results are well understood by clinicians and policy makers. Substandard methods and poor reporting of these studies have limited our ability to do this to date, including having to withdraw tests from the market that have been shown to have poor test accuracy[60,61]. Poor communication about the intended roles and diagnostic performance of tests has led to tests being used inappropriately, for example antibody tests being used to screen or diagnose patients with acute infections[62] or using inaccurate rapid testing to screen asymptomatic individuals and falsely reassuring individuals who are infectious[15]. The issues regarding determining the clinical performance of antibody tests have been particularly challenging.

Inflated and inappropriate claims for test accuracy have been made for tests during the pandemic[63,64]. Most tests have been evaluated by the teams that have developed the tests using convenience samples. More accurate estimates would be derived using prospectively collected samples representing the target population, ideally evaluated by independent teams. This has been a particular problem for the evaluation of antibody tests[8]. Submissions for emergency use authorization should be made publicly available to allow critical review, and data should be made available for use in individual patient data meta-analyses. Leading international and national public health organizations, regulatory authorities and scientific journal editorial boards could assist by harmonizing their requirements for test evaluations, developing study templates that can be used across studies and that encourage standardized data collection and reporting and encourage rigorous study design.

Acknowledgements: The authors thank Dr Tze Ping Loh, Mary Kathryn Bohn and Shannon Steele for their helpful assistance and comments on the manuscript, and Dr Ian Doust for providing Figure 2.

References

1. Lu R, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. 2020;395(10224):565-574. doi:10.1016/S0140-6736(20)30251-8

2. Zhang Y-Z. Novel 2019 coronavirus genome. Virological. Available from: http://virological.org/t/novel-2019-coronavirus-genome/319

3. Foundation for Innovative New Diagnostics website. https://www.finddx.org/covid-19/ [accessed 25 July 2020]

4. JRC Covid-19 In Vitro Diagnostic Devices and Test Methods. https://covid-19-diagnostics.jrc.ec.europa.eu/ [accessed 25 July 2020]

5. Norwegian Institute of Public Health Systematic and Living Map on COVID-19 Evidence https://www.nornesk.no/forskningskart/NIPH_diagnosisMap.html [accessed 20 August 2020]

6. Tromberg BJ, Schwetz TA, Pérez-Stable EJ, Hodes RJ, Woychik RP, Bright RA, Fleurence RL, Collins FS. Rapid Scaling Up of Covid-19 Diagnostic Testing in the United States - The NIH RADx Initiative. N Engl J Med. 2020 Jul 22. doi: 10.1056/NEJMsr2022263. Epub ahead of print. PMID: 32706958.

7. National Institute for Health and Care Excellence. Diagnostic tests for COVID-19 – Evidence Standards Framework. https://www.nice.org.uk/Media/Default/About/what-we-do/covid-19/Diagnostic-tests-for-COVID-19-evidence-standards-framework.pdf. [Accessed July 25 2020]

8. Deeks JJ, Dinnes J, Takwoingi Y, et al. Antibody tests for identification of current and past infection with SARS-CoV-2. Cochrane Database Syst Rev. 2020 Jun 25;6(6):CD013652. doi: 10.1002/14651858.CD013652. PMID: 32584464; PMCID: PMC7387103.

9. Lisboa Bastos M, Tavaziva G, Abidi SK, et al. Diagnostic accuracy of serological tests for covid-19: systematic review and meta-analysis. BMJ. 2020 Jul 1;370:m2516. doi: 10.1136/bmj.m2516. PMID: 32611558; PMCID: PMC7327913.

10. Kontou PI, Braliou GG, Dimou NL, Nikolopoulos G, Bagos PG. Antibody Tests in Detecting SARS-CoV-2 Infection: A Meta-Analysis. Diagnostics (Basel). 2020 May 19;10(5):319. doi: 10.3390/diagnostics10050319. PMID: 32438677; PMCID: PMC7278002.

11. Mallett S, Allen AJ, Graziadio S, Taylor SA, Sakai NS, Green K, Suklan J, Hyde C, Shinkins B, Zhelev Z, Peters J, Turner PJ, Roberts NW, di Ruffano LF, Wolff R, Whiting P, Winter A, Bhatnagar G, Nicholson BD, Halligan S. At what times during infection is SARS-CoV-2 detectable and no longer detectable using RT-PCR-based tests? A systematic review of individual participant data. BMC Med. 2020 Nov 4;18(1):346. doi: 10.1186/s12916-020-01810-8. PMID: 33143712; PMCID: PMC7609379.

12. Arevalo-Rodriguez I, Buitrago-Garcia D, Simancas-Racines D, et al. False negative results of initial RT-PCR assays for COVID-19: a systematic review. medRxiv 2020.04.16.20066787; doi: https://doi.org/10.1101/2020.04.16.20066787. https://www.medrxiv.org/content/10.1101/2020.04.16.20066787v2.article-info

13. Weiss A, Jellingsø M, Sommer MOA. Spatial and temporal dynamics of SARS-CoV-2 in COVID-19 patients: A systematic review and meta-analysis. EBioMedicine. 2020

Jul 22;58:102916. doi: 10.1016/j.ebiom.2020.102916. Epub ahead of print. PMID: 32711256; PMCID: PMC7374142.

14. Dinnes J, Deeks JJ, Adriano A, Berhane S, Davenport C, Dittrich S, Emperador D, Takwoingi Y, Cunningham J, Beese S, Dretzke J, Ferrante di Ruffano L, Harris IM, Price MJ, Taylor-Phillips S, Hooft L, Leeflang MM, Spijker R, Van den Bruel A; Cochrane COVID-19 Diagnostic Test Accuracy Group. Rapid, point-of-care antigen and molecular-based tests for diagnosis of SARS-CoV-2 infection. Cochrane Database Syst Rev. 2020 Aug 26;8:CD013705. doi: 10.1002/14651858.CD013705. PMID: 32845525.

15. Deeks JJ, Raffle AE. Lateral flow tests cannot rule out SARS-CoV-2 infection BMJ 2020; 371:m4787

16. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, de Vet HC, Kressel HY, Rifai N, Golub RM, Altman DG, Hooft L, Korevaar DA, Cohen JF; STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015 Oct 28;351:h5527. doi: 10.1136/bmj.h5527. PMID: 26511519; PMCID: PMC4623764.

17. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011 Oct 18;155(8):529-36. doi: 10.7326/0003-4819-155-8-201110180-00009. PMID: 22007046.

18. Food and Drug Administration https://www.fda.gov/medical-devices/coronavirus-disease-2019-covid-19-emergency-use-authorizations-medical-devices/vitro-diagnostics-euas [Accessed July 25 2020]

19. Medicine and Healthcare products Regulatory Agency. Target-Product-Profile-Antibody-Tests-to-Help-Determine-if-People-Have-Recent-Infection-to-SARS-CoV-2-version-2. https://www.gov.uk/government/publications/how-tests-and-testing-kits-for-coronavirus-covid-19-work/target-product-profile-antibody-tests-to-help-determine-if-people-have-recent-infection-to-sars-cov-2-version-2 [Accessed July 25 2020]

20. World Health Organization. Target product profiles for priority diagnostics to support response to the COVID-19 pandemic v. 1.0. Geneva: WHO. 2020.

21. European Commission. COVID-19: Recommendations for testing strategies 2020. https://ec.europa.eu/info/sites/info/files/covid19_-_eu_recommendations_on_testing_strategies_v2.pdf [accessed 25 july 2020]

22. Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02017R0746-20170505 [Accessed September 26 2020]

23. Global Harmonization Task Force. Clinical Evidence for IVD medical devices – Scientific Validity Determination and Performance Evaluation. http://www.imdrf.org/docs/ghtf/final/sg5/technical-docs/ghtf-sg5-n7-2012-scientific-validity-determination-evaluation-121102.pdf [Accessed September 2020]

24. Infectious Diseases Society of America. Guidelines on the Diagnosis of COVID-19. https://www.idsociety.org/COVID19guidelines/dx [Accessed September 2020]

25. Stephens DS, McElrath MJ. COVID-19 and the Path to Immunity. JAMA. 2020 Sep 11. doi: 10.1001/jama.2020.16656. Epub ahead of print. PMID: 32915201.

26. Wajnberg A, Amanat F, Firpo A et al. SARS-CoV-2 infection induces robust, neutralizing antibody responses that are stable for at least three months. medRxiv 2020.07.14.20151126; doi: https://doi.org/10.1101/2020.07.14.20151126 https://www.medrxiv.org/content/10.1101/2020.07.14.20151126v1

27. Sewell HF, Agius RM, Stewart M, Kendrick D. Cellular immune responses to covid-19. BMJ. 2020 Jul 31;370:m3018. doi: 10.1136/bmj.m3018. PMID: 32737031.

28. Kretzschmar ME, Rozhnova G, Bootsma MCJ, van Boven M, van de Wijgert JHHM, Bonten MJM. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. Lancet Public Health. 2020 Aug;5(8):e452-e459. doi: 10.1016/S2468-2667(20)30157-2. Epub 2020 Jul 16. PMID: 32682487; PMCID: PMC7365652.

29. Woloshin S, Patel N, Kesselheim AS. False Negative Tests for SARS-CoV-2 Infection - Challenges and Implications. N Engl J Med. 2020 Aug 6;383(6):e38. doi: 10.1056/NEJMp2015897. Epub 2020 Jun 5. PMID: 32502334.

30. Lord SJ, St John A, Bossuyt PM, Sandberg S, Monaghan PJ, O'Kane M, Cobbaert CM, Röddiger R, Lennartz L, Gelfi C, Horvath AR; Test Evaluation Working Group of the European Federation of Clinical Chemistry and Laboratory Medicine. Setting clinical performance specifications to develop and evaluate biomarkers for clinical use. Ann Clin Biochem. 2019 Sep;56(5):527-535. doi: 10.1177/0004563219842265. Epub 2019 Apr 16. PMID: 30987429.

31. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ. 2006 May 6;332(7549):1089-92. doi: 10.1136/bmj.332.7549.1089. Erratum in: BMJ. 2006 Jun 10;332(7554):1368. PMID: 16675820; PMCID: PMC1458557.

32. He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, Lau YC, Wong JY, Guan Y, Tan X, Mo X, Chen Y, Liao B, Chen W, Hu F, Zhang Q, Zhong M, Wu Y, Zhao L, Zhang F, Cowling BJ, Li F, Leung GM. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat Med. 2020 May;26(5):672-675. doi: 10.1038/s41591-020-0869-5. Epub 2020 Apr 15. Erratum in: Nat Med. 2020 Aug 7;: PMID: 32296168.

33. Wölfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Müller MA, Niemeyer D, Jones TC, Vollmar P, Rothe C, Hoelscher M, Bleicker T, Brünink S, Schneider J, Ehmann R, Zwirglmaier K, Drosten C, Wendtner Cet al. Virological assessment of hospitalized patients with COVID-2019. Nature. 2020 May;581(7809):465-469. doi: 10.1038/s41586-020-2196-x. Epub 2020 Apr 1. PMID: 32235945.

34. Walsh KA, Jordan K, Clyne B, Rohde D, Drummond L, Byrne P, Ahern S, Carty PG, O'Brien KK, O'Murchu E, O'Neill M, Smith SM, Ryan M, Harrington P. SARS-CoV-2 detection, viral load and infectivity over the course of an infection. J Infect. 2020 Sep;81(3):357-371. doi: 10.1016/j.jinf.2020.06.067. Epub 2020 Jun 29. PMID: 32615199; PMCID: PMC7323671.

35. Lee S, Kim T, Lee E, Lee C, Kim H, Rhee H, Park SY, Son HJ, Yu S, Park JW, Choo EJ, Park S, Loeb M, Kim TH. Clinical Course and Molecular Viral Shedding Among Asymptomatic and Symptomatic Patients With SARS-CoV-2 Infection in a Community Treatment Center in the Republic of Korea. JAMA Intern Med. 2020 Aug

6:e203862. doi: 10.1001/jamainternmed.2020.3862. Epub ahead of print. PMID: 32780793; PMCID: PMC7411944.

36. Gudbjartsson DF, Norddahl GL, Melsted P, et al. Humoral Immune Response to SARS-CoV-2 in Iceland. N Engl J Med. 2020 Sep 1. doi: 10.1056/NEJMoa2026116. Epub ahead of print. PMID: 32871063.

37. Service RF. Fast, cheap tests could enable safer reopening. Science. 2020 Aug 7;369(6504):608-609. doi: 10.1126/science.369.6504.608. PMID: 32764044.

38. Mitchell SL, St George K, Rhoads DD, et al. Understanding, Verifying, and Implementing Emergency Use Authorization Molecular Diagnostics for the Detection of SARS-CoV-2 RNA. J Clin Microbiol. 2020 Jul 23;58(8):e00796-20. doi: 10.1128/JCM.00796-20. PMID: 32381642; PMCID: PMC7383533.

39. Theel E, Filkins L, PalavecinoE, et al. Verification procedure for commercial serologic tests with Emergency Use Authorization for detection of antibodies to SARS-CoV-2. American Society for Microbiology. 24 June 2020. https://asm.org/Protocols/Verify-Emergency-Use-Authorization-EUA-SARS-CoV-2. [Accessed July 25 2020]

40. Willman D. Contamination at CDC Lab Delayed Rollout of Coronavirus Tests. Washington Post. 2020, April 19. https://www.washingtonpost.com/investigations/contamination-at-cdc-lab-delayed-rollout-of-coronavirus-tests/2020/04/18/fd7d3824-7139-11ea-aa80-c2470c6b2034_story.html. [Accessed July 25 2020]

41. Clinical and Laboratory Standards Institute. Harmonized Terminology Database. https://htd.clsi.org/listterms.asp?searchd. [accessed July 25 2020]

42. Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. J Clin Epidemiol. 2006 Aug;59(8):798-801. doi: 10.1016/j.jclinepi.2005.11.025. Epub 2006 May 26. Erratum in: J Clin Epidemiol. 2007 Jul;60(7):756. PMID: 16828672.

43. Eyre DW, Lumley SF, O'Donnell D et al. Stringent thresholds for SARS-CoV-2 IgG assays result in underdetection of cases reporting loss of taste/smell. https://www.medrxiv.org/content/10.1101/2020.07.21.20159038v1

44. Bossuyt PM. Testing COVID-19 tests faces methodological challenges. J Clin Epidemiol. 2020 Jul 3:S0895-4356(20)30750-2. doi: 10.1016/j.jclinepi.2020.06.037. Epub ahead of print. PMID: 32622902; PMCID: PMC7332449.

45. World Health Organization. Global surveillance for COVID-19 caused by human infection with COVID-19 virus, interim guidance. Issued 20 March 2020

46. World Health Organization. Laboratory testing of 2019 novel coronavirus (2019-nCoV) in suspected human cases: interim guidance. Issued 21st March 2020

47. Long C, Xu H, Shen Q, et al. Diagnosis of the Coronavirus disease (COVID-19): rRT-PCR or CT? Eur J Radiol. 2020 May;126:108961. doi: 10.1016/j.ejrad.2020.108961. Epub 2020 Mar 25. PMID: 32229322; PMCID: PMC7102545.

48. Kucirka LM, Lauer SA, Laeyendecker O, et al. Variation in False-Negative Rate of Reverse Transcriptase Polymerase Chain Reaction-Based SARS-CoV-2 Tests by Time Since Exposure. Ann Intern Med. 2020 May 13:M20-1495. doi: 10.7326/M20-1495. Epub ahead of print. PMID: 32422057; PMCID: PMC7240870.

49. Pan Y, Zhang D, Yang P, et al. Viral load of SARS-CoV-2 in clinical samples. Lancet Infect Dis. 2020 Apr;20(4):411-412. doi: 10.1016/S1473-3099(20)30113-4. Epub 2020 Feb 24. PMID: 32105638; PMCID: PMC7128099.

50. Zhao Y, Xia Z, Liang W, et al. SARS-CoV-2 persisted in lung tissue despite disappearance in other clinical samples. Clin Microbiol Infect. 2020 May 22:S1198-743X(20)30290-1. doi: 10.1016/j.cmi.2020.05.013. Epub ahead of print. PMID: 32447048; PMCID: PMC7242209.

51. Zou MY, Wu GQ. The effect of antigen cross-reaction on testing of SARS-CoV-2 specific antibodies in serum. Chinese Journal of Clinical Laboratory Science 2020;3:161-3.

52. Zhang R, Li JM. Talking about false positive testing result of SARS-CoV-2 specific antibodies (IgM/IgG). National Center for Clinical Laboratories, 2020.

53. Hadgu A. Discrepant analysis: a biased and an unscientific method for estimating test sensitivity and specificity. J Clin Epidemiol. 1999 Dec;52(12):1231-7. doi: 10.1016/s0895-4356(99)00101-8. PMID: 10580787.

54. Bullard J, Dust K, Funk D, Strong JE, Alexander D, Garnett L, Boodman C, Bello A, Hedley A, Schiffman Z, Doan K. Predicting infectious SARS-CoV-2 from diagnostic samples. Clinical Infectious Diseases. 2020 May 22.

55. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med. 1998 Apr 30;17(8):857-72. doi: 10.1002/(sici)1097-0258(19980430)17:8<857::aid-sim777>3.0.co;2-e. PMID: 9595616.

56. Wilson, E. B. Probable Inference, the Law of Succession, and Statistical Inference. Journal of the American Statistical Association. 1927: 22, 209-212.

57. Glasziou P, Irwig L, Deeks JJ. When should a new test become the current reference standard? Ann Intern Med. 2008 Dec 2;149(11):816-22. doi: 10.7326/0003-4819-149-11-200812020-00009. PMID: 19047029.

58. Food and Drug Administration. https://www.fda.gov/medical-devices/emergency-situations-medical-devices/eua-authorized-serology-test-performance

59. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, Irwig L, Levine D, Reitsma JB, de Vet HC, Bossuyt PM. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open. 2016 Nov 14;6(11):e012799. doi: 10.1136/bmjopen-2016-012799. PMID: 28137831; PMCID: PMC5128957.

60. Loh T. New Tests Could Turn Tide Against Coronavirus if They Work. Bloomberg. 2020, March 31. https://www.bloomberg.com/news/articles/2020-03-31/new-tests-could-turn-tide-against-coronavirus-if-they-work

61. Hagemann H. Antibody Tests Go to Market Largely Unregulated Warns House Subcommittee Chair. NPR. 2020, April 25. https://www.npr.org/sections/coronavirus-live-updates/2020/04/26/845164212/antibody-tests-go-to-market-largely-unregulated-warns-house-subcommittee-chair

62. Calafiore S. GPs face $20k fines for using serology tests to diagnose coronavirus. https://www.rcpa.edu.au/Library/COVID-19-Updates/COVID-19-Useful-Resources/Docs/GPs-face-$20k-fines-for-using-serology-tests-to-di.aspx

63. Gross A, Kelly J. Is the company with a 20-second coronavirus test for real? *Financial Times* 2020 Sep 16, https://www.ft.com/content/e7a279df-3239-4e00-be29-f38d98f4d730.

64. Bosely, S. Claims of 99% accuracy for UK Covid antibody test ‘cannot be trusted https://www.theguardian.com/world/2020/aug/27/data-secrecy-covid-antibody-test-trusted-fingerprint-doubt

65. Centers for Disease Control. Interim Guidance for Antigen Testing for SARS-CoV-2. 2020 December 16. https://www.cdc.gov/coronavirus/2019-ncov/lab/resources/antigen-tests-guidelines.html.

66. Centers for Disease Control. Interim Guidance for COVID-19 Antibody testing. 2020 August 1. https://www.cdc.gov/coronavirus/2019-ncov/lab/resources/antibody-tests-guidelines.html

Box 1: Terminology used in this guidance

Clinical performance studies: assess the ability of a test to discriminate those who have the target condition from those who do not have the target condition in clinical or public health practice[16].

Scientific validity studies: establish an association between an analyte and a clinical condition or physiological state[20]. For SARS-CoV-2 tests, they are often performed on artificial or restricted sample sets, for example comparing residual samples from individuals hospitalised with COVID-19 (cases) with pre-2020 samples (controls).

Analytical performance studies: refers to technical test performance, and may include data to demonstrate accuracy (derived from trueness and precision), analytical sensitivity (eg limit of detection, limit of quantitation), analytical specificity, linearity, cut-off, measuring interval (range), carry-over, as well as determination of appropriate specimen collection and handling, and endogenous and exogenous interference on assay results[21].

Target condition: a particular disease, disease stage, health status, or any other identifiable condition within a patient, such as staging a disease already known to be present, or a health condition that should prompt clinical action, such as the initiation, modification, or termination of treatment[16]

Index test: the test being evaluated[16].

Reference standard: the best available method for establishing the presence or absence of the target condition related to the intended use of the test[16].

Reference method: used in analytical studies to refer to the best analytical method to detect a measurand.

Reverse Transcription Polymerase Chain Reaction (RT-PCR): A molecular test using cyclical amplification of DNA to detect if genetic material consistent with the SARS-CoV-2 virus is present in the sample (through a DNA mold, that is the reverse transcription of the viral RNA).

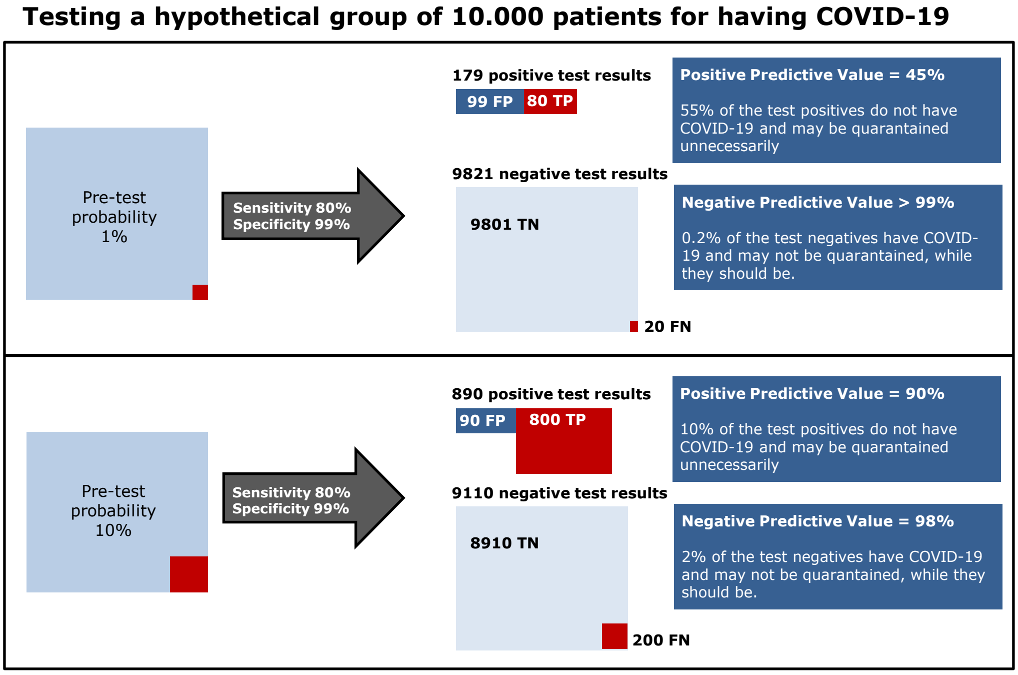
Cycle threshold (C_T): Each cycle of RT-PCR amplifies the number of DNA copies in the sample. The more virus that is present the less amplification is needed to detect the virus. Laboratories will run samples through machines with a set numbers of cycles (typically 40 to 50 cycles), and will establish a threshold for when a sample is determined to be positive, for example 35 or 40. Samples that test positive after this threshold may be retested.

Antigen testing: immunoassays that detect the presence of a specific viral antigen, which implies current viral infection[65].

Lateral flow test: A form of immunoassay performed outside of the laboratory using a sample placed onto a test device, with the presence or absence of the target analyte demonstrated by a colour change. A common example is a pregnancy test. In this context, they are used to detect SARS-CoV-2 antigens or antibodies.

Antibody testing: Serologic or antibody tests detect resolving or past SARS-CoV-2 virus infection indirectly by measuring the person's humoral immune response to the virus[66].

Figure 1: Graphical demonstration of the positive and negative predictive values of testing, based on the pre-test probability and sensitivity and specificity of testing.


Testing a hypothetical group of 10.000 patients for having COVID-19
Pre-test probability 1%
Sensitivity 80%
Specificity 99%
179 positive test results
99 FP
80 TP
Positive Predictive Value = 45%
55% of the test positives do not have COVID-19 and may be quarantined unnecessarily
9821 negative test results
9801 TN
20 FN
Negative Predictive Value > 99%
0.2% of the test negatives have COVID-19 and may not be quarantained, while they should be.
Pre-test probability 10%
Sensitivity 80%
Specificity 99%
890 positive test results
90 FP
800 TP
Positive Predictive Value = 90%
10% of the test positives do not have COVID-19 and may be quarantained unnecessarily
9110 negative test results
8910 TN
200 FN
Negative Predictive Value = 98%
2% of the test negatives have COVID-19 and may not be quarantained, while they should be.

FP: false positive;
TP: true positive;
TN: true negative;
FN: false negative;
Sensitivity: the proportion of participants with the target condition who have a positive index test;
Specificity: the proportion without the target condition who have a negative index test
Positive predictive value: the proportion of participants with a positive index test who have the target condition;
Negative predictive value: the proportion of participants with a negative index test who do not have the target condition

Figure 2: Schematic diagram of the timing of tests for SARS-CoV-2[8,32-34]

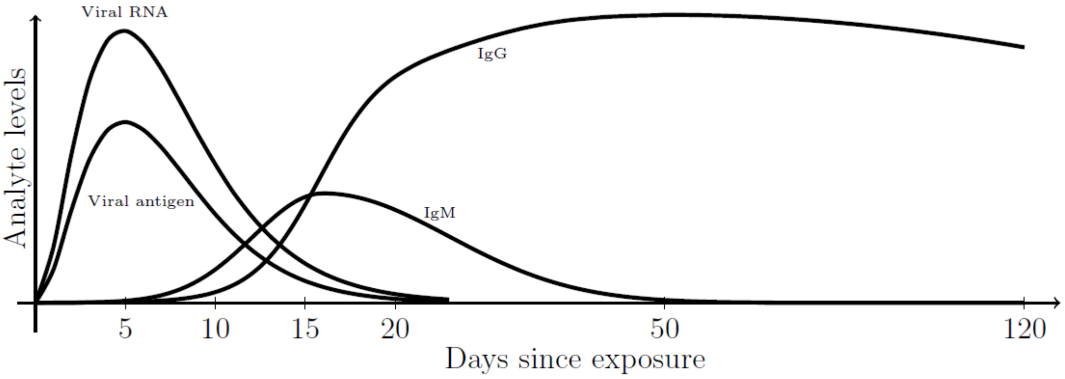


Figure 3: Prototypical flow diagram for participants in studies evaluating diagnostic accuracy[16]

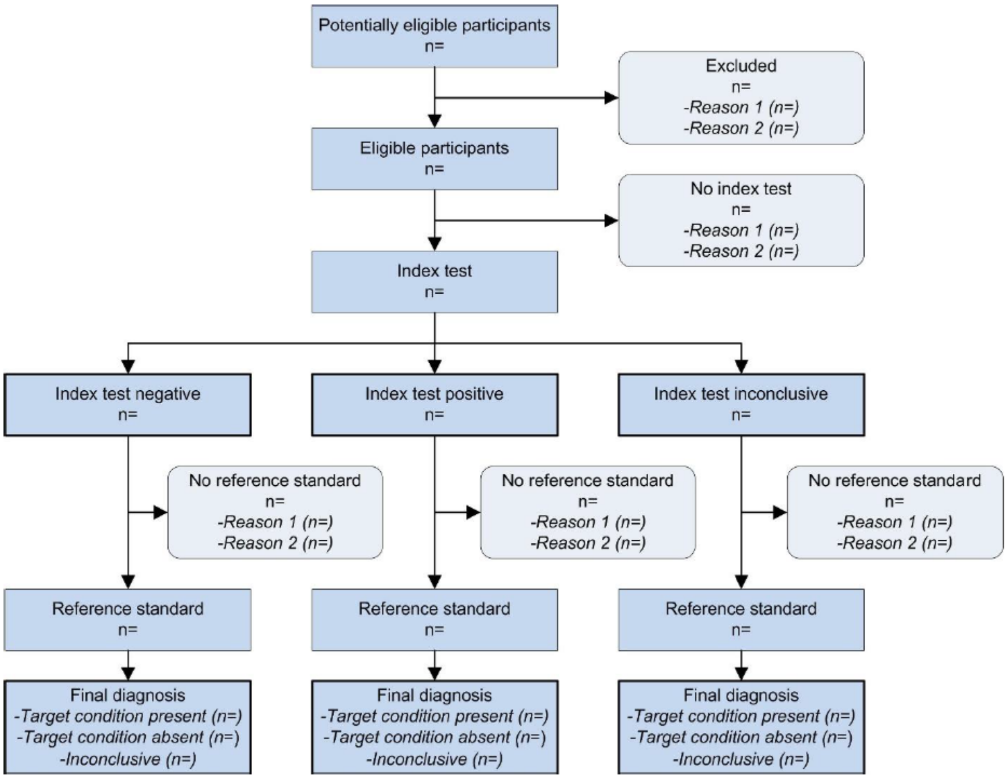


Table 1: The STARD checklist[16] and problems noted in studies of SARS-CoV-2 clinical performance studies[8-14]

Section & Topic	No	Item	Step in this guidance	Problems noted in studies of SARS-CoV-2 tests to date
TITLE OR ABSTRACT				
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	1	Diagnostic accuracy results reported but are not included as a study objective (for example in seroprevalence studies or studies of antibody patterns).
ABSTRACT				
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	7	Study design labels not clear. Preprints often do not include abstracts.
INTRODUCTION				
	3	Scientific and clinical background, including the intended use and clinical role of the index test	1,2	Lack of clarity of the intended use and target condition, for example whether the target condition is the presence of the virus, infectivity, or presence of COVID-19. Scientific validity studies (eg case-control studies) being used inappropriately to estimate clinical performance.
	4	Study objectives and hypotheses	1	Not establishing if the objective of the study is to establish scientific validity or clinical performance/diagnostic accuracy. Not stating if clinical performance is a study objective.
METHODS				
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	4,5	Not reporting when the data were collected, especially when healthy control samples used. Enrolling patients in studies based on PCR test results.
Participants	**6**	Eligibility criteria	3	Not reporting or recording the symptoms or other features used to enrol patients in the study. Not reporting the time of either the index test or the reference standard in relation to key clinical time points, such as time since a high-risk contact or onset of symptoms.
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	3	Including participants hospitalised with COVID-19 to establish the sensitivity of a test. Including pre-COVID-19 banked specimens to establish the specificity of a test. Excluding patients with other respiratory illnesses
	8	Where and when potentially eligible participants were identified (setting, location and dates)	3	Not being clear what hospital departments were involved for studies done in a hospital. Using samples submitted for routine laboratory testing but not stating when or where samples were submitted from.
	9	Whether participants formed a consecutive, random or convenience series	3	Not enrolling a consecutive series of patients aimed at a clinical use, for example patients suspected of having SARS-CoV-2 infection.
Test methods	**10a**	Index test, in sufficient detail to allow replication	4	Not reporting the anatomical site used for the collection of the specimen. Not reporting who obtained the sample or who carried out and interpreted the test. No details of product codes for commercially available tests. Using viral transport medium spiked with inactivated virus
	10b	Reference standard, in sufficient detail to allow replication	6	Reference standard often reported in insufficient detail to allow replication, often using in-house unpublished methods with unclear analytical and clinical performance.
	11	Rationale for choosing the reference standard (if alternatives exist)	6	Difficulty in applying the reference standard, for example using the WHO case definition of COVID-19

	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	4	Distinction between cut-offs that are pre-specified or exploratory is often not made.
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	4	Distinction between cut-offs that are pre-specified or exploratory is often not made. Threshold for positivity and how this was determined often not reported
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	6	Information available to the assessors of the index test not reported. Not possible to determine which test was carried out first (and therefore blinded)
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	6	Information available to the assessors of the reference standard not reported
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	7	Calculation of sensitivity and specificity rarely explained, as well as how the categories of those with and without the target condition were defined.
	15	How indeterminate index test or reference standard results were handled	7	Often not reported. Flow diagrams demonstrating indeterminant results not included
	16	How missing data on the index test and reference standard were handled	7	Rarely reported; studies often only report positive and negative tests, with intermediate test results and test failures excluded and/or not documented
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	7	Often not reported
	18	Intended sample size and how it was determined	7	Sample size estimations require information about the expected or target accuracy of the index test, which is often not reported.
RESULTS				
Participants	**19**	Flow of participants, using a diagram	7	Few studies provide flow charts demonstrating the flow of participants, including timing, indeterminate and missing results
	20	Baseline demographic and clinical characteristics of participants	7	Demographics and baseline clinical characteristics are often not reported.
	21a	Distribution of severity of disease in those with the target condition	7	Severity definitions and distributions rarely provided, prevalence of infection often not reported
	21b	Distribution of alternative diagnoses in those without the target condition		Alternative diagnoses may sometimes be part of the reference standard to indicate someone as not having SARS-CoV-2, although co-infections do not preclude SARS-CoV-2 infection.
	22	Time interval and any clinical interventions between index test and reference standard	7	This is usually not an issue, as index test and reference standard are done at the same time, e.g. using sample or paired samples. Some examples of 6-24h delays in paired sample collection.
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	7	Cross-tabulation of results (a 2x2 table) not provided
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	7	Confidence intervals are sometimes not reported
	25	Any adverse events from performing the index test or the reference standard		Direct adverse events are not applicable in most situations of SARS-CoV-2 tests
DISCUSSION				
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	7	Assuming that the results seen in a reference laboratory or a clinical setting with patients with a high prevalence of infection will be achieved in other clinical settings

	27	Implications for practice, including the intended use and clinical role of the index test	7	The role of the index test is rarely explained, although can sometimes be deduced from the study design. Overstatement of implications from results in terms of significance for practice or assuming generalisability to other settings.
OTHER INFORMATION				
	28	Registration number and name of registry	8	Rarely reported. Clinical performance studies often not pre-registered
	29	Where the full study protocol can be accessed	8	Rarely reported
	30	Sources of funding and other support; role of funders	8	Co-author affiliation to commercial manufacturers may only be derived from author institutions rather than COI statements, Regulatory status of producer often not reported

Table 2: Examples of possible study designs to evaluate the clinical performance of SARS-CoV-2 tests used for different purposes

	Purpose of testing				
	Diagnosis	**Test and trace programs**	**Determining if an individual is infectious**	**Assessing seroprevalence**	**Assessing protective immune response from vaccination**
Intended use of test	To diagnose COVID-19 in individuals with symptoms suggestive of the disease	To screen individuals exposed to confirmed cases of SARS-CoV-2 in test-and-trace programs for infection	To rapidly determine if an individual is infectious, for example in a healthcare setting	To estimate seroprevalence in a population as a measure of exposure to SARS-CoV-2 infection	To evaluate if a vaccine has generated protective immunity
Target condition	COVID-19	Current SARS-CoV-2 infection	SARS-CoV-2 infectivity	Recent and past SARS-CoV-2 infection	Protective immunity to SARS-Cov-2
Minimal clinical performance characteristics	Emphasis on high sensitivity to reduce the risk of missed disease (false negatives)	Emphasis on high sensitivity to reduce the risk of missed case of infection (false negatives)	Lower specificity may be acceptable if positive results are confirmed with later testing	Emphasis on high specificity to reduce the potential for false positives to account for all/most positive results in populations where prevalence is low[21]	Emphasis on high specificity to reduce the potential for people thought to have immunity when they do not (false positives)
Study population	Symptomatic individuals in community and/or in hospital	Asymptomatic, pre-symptomatic or mildly symptomatic individuals in the community	Individuals presenting in a health care setting	Randomly selected pre-symptomatic or asymptomatic individuals from a population potentially exposed to SARS-CoV-2 virus	Individuals who received SARS-Cov-2-specific vaccine
Index test	RT-PCR test (e.g. naso-pharyngeal swab)	RT-PCR test (e.g. naso-pharyngeal swab)	Point of care test (e.g. RT-LAMP test on nasal swab or saliva)	Antibody test (e.g serum)	Antibody test that detects antibodies with virus neutralizing capacity (plasma or serum)
Comparator test	-	-	RT-PCR	-	-

Reference standard[a]	Composite of clinical information including specified symptoms and results of tests such as RT-PCR, antigen testing, chest imaging and clinical follow-up	Composite to determine presence/absence of current infection, e.g. repeated RT-PCR and epidemiological information such as exposure risk	Measure of infectivity - Acceptable reference does not currently exist	Composite to determine presence/absence of recent/past infection, e.g. repeated RT-PCR and epidemiological information such as exposure risk	Measures of the overall humoral and cellular immune response to SARS-CoV-2 vaccine
Timing of index test	First 2 weeks after symptom onset	First 2 weeks after symptom onset or exposure	Representative of target population (with timing of exposure/infection recorded if known)	> 2 weeks after exposure for those where infection is established	> 2 weeks after vaccination
Other possible outcomes/ considerations	Turnaround time, burden on laboratories and personnel, ability to use outside of a medical setting, potential infectivity of samples				

RT-PCR: Reverse transcription polymerase chain reaction; POCT: Point of care test; RT-LAMP: Reverse transcription loop-mediated isothermal amplification

a: Note that all reference standards described here are not infallible. For example, the use of a composite reference standard using all clinical information will incorporate the index test so will give biased estimates of diagnostic accuracy.

No reference standard that detects both humoral and cellular immunity is currently available. Reference standards defining humoral immunity by capturing seroconversion are not a surrogate for overall immune response, and the presence or absence of even neutralizing antibodies does not rule in or out protective immunity. New data from vaccine trials are needed to define what study design and reference standard would best test for immunity following vaccination.