FAST TRACK

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

Jenny A Doust,[1] Katy J L Bell,[2] Mariska M G Leeflang,[3] Jacqueline Dinnes,[4,5] Sally J Lord,[6] Sue Mallett,[7] Janneke H H M van de Wijgert,[8,9] Sverre Sandberg,[10,11] Khosrow Adeli,[12,13] Jonathan J Deeks,[4,5] Patrick M Bossuyt,[3] Andrea R Horvath[2,14,15]

For numbered affiliations see end of the article.

Correspondence to: J Doust j.doust@uq.edu.au (ORCID 0000-0002-4024-9308)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* **2021;372:n568**
http://dx.doi.org/10.1136/bmj.n568

Accepted: 22 February 2021

Testing for SARS-CoV-2 infection is key in managing the current pandemic. 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 these studies have identified many methodological issues, leading to a high risk of bias and difficulties applying the results in practice. Better guidance is urgently needed on the conduct and interpretation of these studies. This article outlines the principles for defining the intended purpose of the test; study population selection; reference standard, test timing; and other critical considerations for the design, reporting, and interpretation of diagnostic accuracy studies. The implementation and accuracy of SARS-CoV-2 tests have major implications for individuals and communities, balancing the potential consequences of continued infection against the need for public health measures, such as the restriction of movements and social activities. Decision making in the current pandemic requires a clear understanding of the clinical performance and limitations of testing. This article provides guidance to assist researchers design robust diagnostic accuracy studies, assist publishers and peer reviewers to assess such studies, and support clinicians and policy makers in their evaluation of the evidence on SARS-CoV-2 testing for clinical and public health decisions. The guidance aims to ensure that studies evaluating the diagnostic accuracy of SARS-CoV-2 tests are conducted as rigorously as possible, in an efficient and timely way.

SUMMARY POINTS

Criticisms of current implementation trials include risks of bias, lack of theory use, lack of standardised terminology to describe implementation strategies, and limited measures and poor reporting

This article consolidates recent methodological developments in implementation science with established guidance from seminal texts of randomised trial methods to provide best practice guidance to improve the development and conduct of randomised implementation trials

Consideration of such guidance will improve the quality and use of randomised implementation trials for healthcare and public health improvement

Testing for infection has a critical role in the response to the pandemic caused by 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 development 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[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 January 2021.[5] The volume of available evaluations of diagnostic test accuracy is unprecedented and is unlikely to diminish with the implementation of programmes to accelerate the development of new tests, such as the Rapid Acceleration of Diagnostics programme 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).[9-15] These problems limit the ability of clinicians and policy makers to apply the results of such studies in diagnostic pathways and public health programmes and have led to poor clinical and public health decisions contributing to ongoing spread of the infection.[16]

This article aims to outline general principles for studies that evaluate the clinical performance of SARS-CoV-2 tests. Here, we use the term "SARS-CoV-2 tests" to refer to any of the following: viral nucleic acid, antigen, antibody, or other 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,[8] 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 US 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 recommendations for covid-19 testing strategies[21] and related documents.[22]

The article focuses on clinical performance studies investigating 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 (box 1).[25] They are essential in the development of a test, analogous to the finding of phase I 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 might 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 relation 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 before the conduct of the study. Before beginning an evaluation of a SARS-CoV-2 test, researchers should 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 listed below.[26]

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, presymptomatic individuals, or individuals with mild symptoms who have known recent exposure to another person with confirmed covid-19 (eg, as part of localised outbreak investigations and test and trace programmes)
3. To screen individuals at risk of acquisition or transmission of infection (eg, staff or patients in hospital or staff or residents in aged care or education facilities, as part of outbreak prevention programmes)
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 (eg, 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 (eg, 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 (eg, for public health decisions).

Summary points

- Many evaluations of SARS-CoV-2 tests do not provide accurate estimates of the performance of the tests in the intended clinical setting
- Studies need to clarify if they are scientific validity studies or clinical performance studies
- The purpose of the test, the study population, the methods for determining the decision thresholds for the test being evaluated, and the reference standard need to be carefully mapped out in the study design
- Studies that compare diagnostic tests and diagnostic pathways, preferably by investigators independent of those who developed the tests, are valuable

Table 1 | STARD checklist[8] and problems noted in studies of SARS-CoV-2 clinical performance studies[9-15]

Section and 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 (eg, 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, infectiousness, 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 or 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 admitted to hospital 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-off thresholds or result categories of the index test, distinguishing prespecified from exploratory	4	Distinction between cut-off thresholds that are prespecified or exploratory is often not made
	12b	Definition of and rationale for test positivity cut-off thresholds or result categories of the reference standard, distinguishing prespecified from exploratory	4	Distinction between cut-off thresholds that are prespecified 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 or 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 indeterminate 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 or not documented
	17	Any analyses of variability in diagnostic accuracy, distinguishing prespecified 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 flowcharts 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

(*Continued*)

Table 1 | Continued

Section and topic	No	Item	Step in this guidance	Problems noted in studies of SARS-CoV-2 tests to date
	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	7	Alternative diagnoses can 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, for example, using sample or paired samples. Some examples of 6-24 hour 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 2×2 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	Coauthor affiliation to commercial manufacturers might only be derived from author institutions rather than COI statements, Regulatory status of producer often not reported

AUC=area under the receiver operating characteristic curve; COI=conflicts of interest; PCR=polymerase chain reaction.

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, or other cellular responses, lead to protection from clinically important infection or re-infection. The detection of antibodies in itself 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.[27-29]

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 turnaround 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 programmes,[30] so in some settings there may be a trade-off between turnaround time and diagnostic accuracy.

False negative test results could 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 (that is, the prevalence in the group) will affect the predictive values of the test (that is, the probability of false positive and false negative test results; fig 1). For example, in settings where there is a very high rate of transmission, the pre-test probability of infection for an individual might 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.[31]

Groups such as the FDA in the US,[18] the MHRA in the UK,[19] and WHO[20] have set acceptable and desirable performance characteristics for SARS-CoV-2 testing (called target product profiles by the MHRA and WHO). 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 before conducting a clinical performance study.[32] 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

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.[8]

Scientific validity studies: establish an association between an analyte and a clinical condition or physiological state.[20] SARS-CoV-2 tests are often performed on artificial or restricted sample sets, for example, comparing residual samples from individuals admitted to hospital with covid-19 with control samples before 2020.

Analytical performance studies: refers to technical test performance, and can include data to demonstrate accuracy (derived from trueness and precision), analytical sensitivity (eg, limit of detection, limit of quantitation), analytical specificity, linearity, cut-off thresholds, measuring interval, cross contamination, 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.[8]

Index test: the test being evaluated.[8]

Reference standard: the best available method for establishing the presence or absence of the target condition related to the intended use of the test.[8]

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 40-50 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 could be retested.

Antigen testing: immunoassays that detect the presence of a specific viral antigen, which implies current viral infection.[23]

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: serological or antibody tests detect resolving or past SARS-CoV-2 virus infection indirectly by measuring the person's humoral immune response to the virus.[24]

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.[33] For example, a reverse transcription loop-mediated isothermal amplification test might be used as a replacement diagnostic test for RT-PCR, to reduce the demand for reagents and allow for faster turnaround 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 third or fourth day after infection (fig 2).[34 35] Symptoms typically appear around the fifth day of infection, and both symptoms and viral detection last for several days to weeks, depending on the severity of infection.[36] Studies using repeat RT-PCR testing and tracking of transmission rates (including infector-infectee transmission pairs) have shown about 40% of transmissions occur before the development of symptoms,[37] and peak infectiousness occurs about one day before until two to three days after symptom onset in typical individuals.[34] Antibodies are generally low in the first week after symptom onset (in people with covid-19 confirmed by 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,[9-11] and then begins to decline.[38] 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.[28 29 38]

Researchers might not be able to predict all aspects of intended uses of the test as well as consequences of the test result. However, researchers should 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, researchers must 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), infectiousness, 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 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, presymptomatic, and asymptomatic infection).

Some settings could require researchers to establish whether someone is infectious rather than whether someone has the infection. For example, if an individual presents in a healthcare setting, knowing whether they are infectious or not influences the need for personal protective equipment and other infection control measures immediately; whereas determining whether they have the infection is less urgent if the individual's symptoms are mild but SARS-CoV-2 infection cannot be excluded. Testing for infectiousness, rather than infection, has also been suggested as a possible

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 programmes	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 person with confirmed SARS-CoV-2 in test and trace programmes	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 infectiousness	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 infection (false negatives)	Lower specificity might 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 or 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 or in hospital	Asymptomatic individuals, presymptomatic individuals, or individuals with mild symptoms in the community	Individuals presenting in a healthcare setting	Randomly selected presymptomatic 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 (eg, nasopharyngeal swab)	RT-PCR test (eg, nasopharyngeal swab)	Point of care test (eg, RT-LAMP test on nasal swab or saliva)	Antibody test (eg, serum)	Antibody test that detects antibodies with virus neutralising capacity (plasma or serum)
Comparator test	—	—	RT-PCR	—	—
Reference standard*	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 or absence of current infection, for example, repeated RT-PCR and epidemiological information such as exposure risk	Measure of infectiousness—acceptable reference does not currently exist	Composite to determine presence or absence of recent or past infection, for example, 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 or infection recorded if known)	>2 weeks after exposure for those where infection is established	>2 weeks after vaccination
Other possible outcomes or 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.
*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 neutralising 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.

method for screening in other settings, including opening businesses and allowing public gatherings.[39] Although 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 people with confirmed disease); individuals at high risk of both exposure and transmission (such as healthcare workers or residents of aged care facilities) and patients admitted to hospital 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 previous 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 people with known disease and healthy controls introduce selection bias and effects related to the clinical spectrum of disease.

The diagnostic accuracy observed in studies of patients admitted to hospital with severe covid-19 or recruited from hospital settings might not apply 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 admitted to hospital with severe infection. Antibody production in this population is likely to be higher than in the wider population of those infected.[9]

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 because these patients will be tested in clinical practice. Careful thought should be given to the presence or absence of symptoms that might 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 (eg, risk of exposure, and time since exposure, if known) and more than one sample (anatomical or time point) might need to be tested. Viral nucleic acid typically can be detected on the third day after exposure in nasal, throat, or saliva secretions.[34 35] It is unclear whether virus is typically detected in faeces and sputum two days after infection, or if later time points are relevant for these sites of sampling.

In addition to defining the population, researchers should 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.

The accuracy of all tests depends on their timing, so 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. Owing to differences in healthcare provision and pathways, only recording time since healthcare events (such as admission to hospital, intensive care units, 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, or tests at different time points might be needed to identify all true cases and non-cases. The index test strategy could therefore be one test, 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 can be developed commercially or in-house by a laboratory, and need to meet key regulatory or emergency use authorisation requirements for in vitro medical devices.[18-22] All pre-analytical, analytical, and postanalytical characteristics of the test should be described, including the items in the list below.

- Full name of the test and manufacturer, and associated batch numbers allowing clear identification

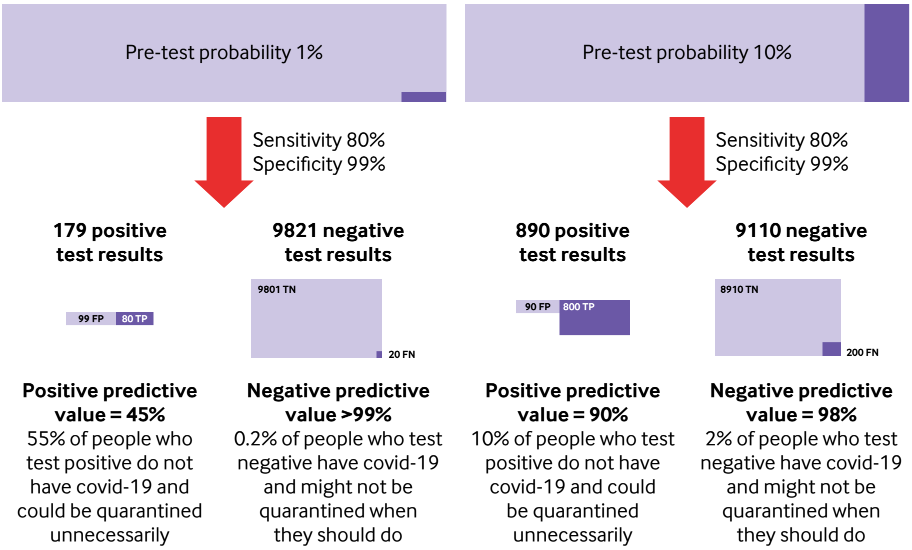
Pre-test probability 1%
Sensitivity 80%
Specificity 99%
179 positive test results
9821 negative test results
99 FP
80 TP
9801 TN
20 FN
Positive predictive value = 45%
55% of people who test positive do not have covid-19 and could be quarantined unnecessarily
Negative predictive value >99%
0.2% of people who test negative have covid-19 and might not be quarantined when they should do
Pre-test probability 10%
Sensitivity 80%
Specificity 99%
890 positive test results
9110 negative test results
90 FP
800 TP
8910 TN
200 FN
Positive predictive value = 90%
10% of people who test positive do not have covid-19 and could be quarantined unnecessarily
Negative predictive value = 98%
2% of people who test negative have covid-19 and might not be quarantined when they should do

Fig 1 | Positive and negative predictive values of testing, based on the pre-test probability and sensitivity and specificity of testing. FP=false positive; TP=true positive; TN=true negative; FN=false negative; sensitivity=proportion of participants with the target condition who have a positive index test; specificity=proportion without the target condition who have a negative index test; positive predictive value=proportion of participants with a positive index test who have the target condition; negative predictive value=proportion of participants with a negative index test who do not have the target condition

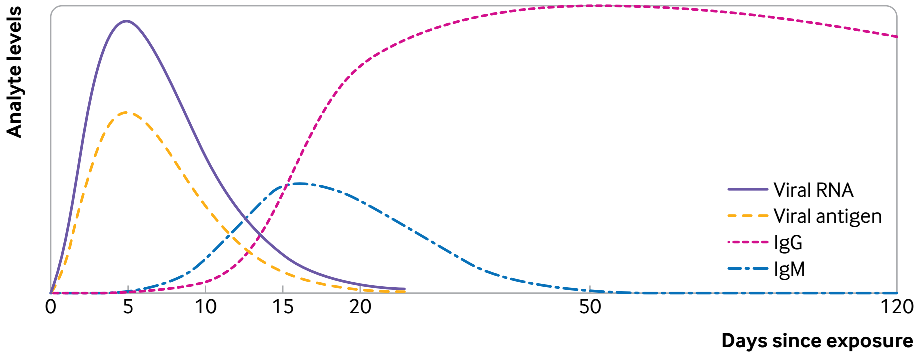

Fig 2 | Timing of tests for SARS-CoV-2[9 34-36]

- Pre-analytical characteristics:
  - type of samples suitable for testing (eg, nasopharyngeal swab, sputum, saliva, blood)
  - method of collection of specimens and how the sample was taken (eg, whether a long swab was used for RT-PCR tests)
  - who has taken the sample (eg, clinical training)
  - conditions for specimen handling, transport, and storage
- Analytical characteristics
  - actual target of the assay (what is being measured; eg, viral nucleic acid, antigen, or antibody against specific viral proteins)
  - principles of analytical methods (eg, fluorescence, multiplex fluorescence, or digital RT-PCR; enzyme linked immunoassay or lateral flow assay)
  - platform used for measurement (how and with what device the target analyte is measured)
  - where was the analysis done, if relevant (eg, at the point of care or in a reference laboratory)
  - analytical performance measures of the test (eg, analytical sensitivity or limit of detection, cross reactivity, accuracy, trueness, precision)
- Postanalytical 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), the lower respiratory tract (bronchoalveolar lavage, sputum), or elsewhere (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 (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, or the nucleocapsid protein, which is conserved among all coronaviruses), the type of immunoglobulin(s) detected (that is, IgA, IgG, or IgM), and the immunological method used (eg, enzyme linked immunosorbent assay, chemiluminescence immunoassays, lateral flow immunoassays, and fluorescent immunoassays). 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 that they adequately measure the presence 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.[40 41] 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

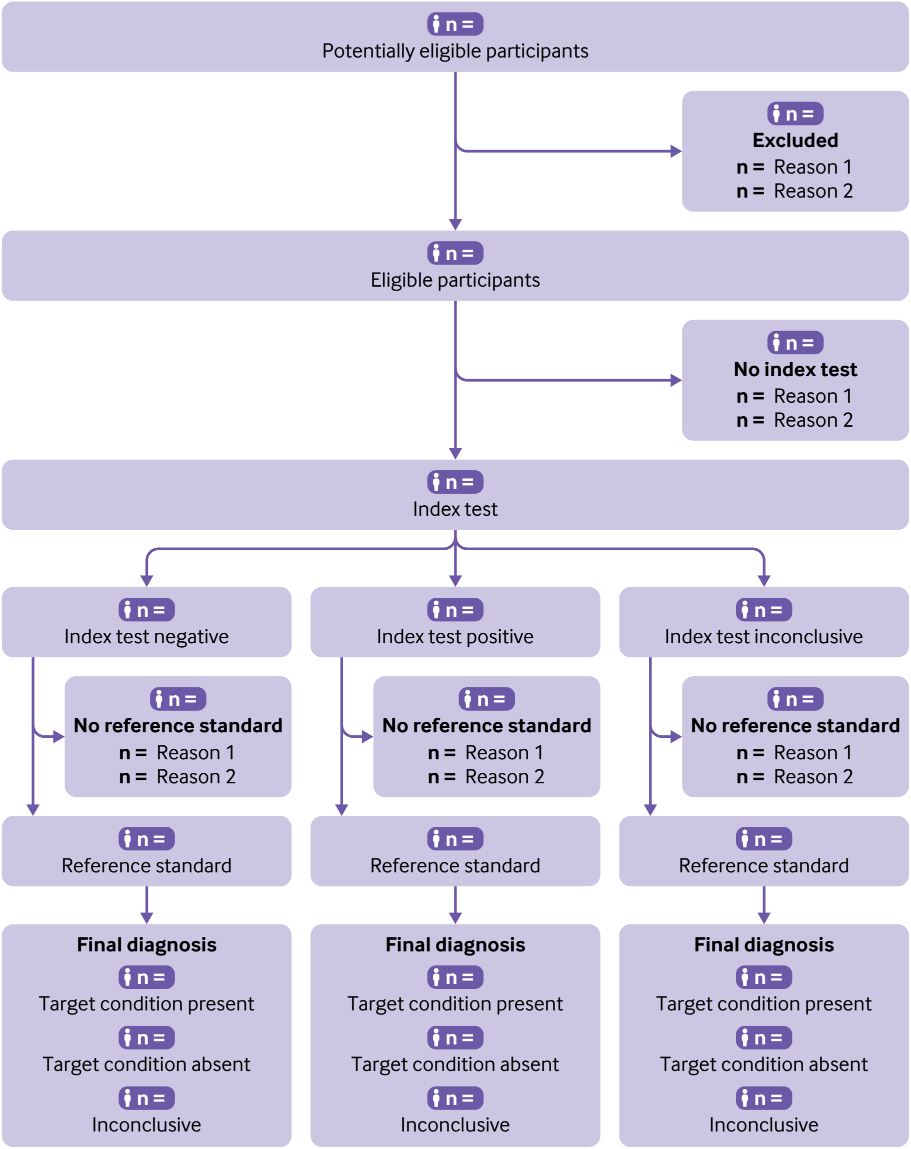

Fig 3 | Prototypical flow diagram for participants in studies evaluating diagnostic accuracy[8]

or real clinical matrix, such as bronchoalveolar lavage fluid or sputum. The limit of detection should be reported, for example, as viral copies per millilitre.

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 might 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, and controlled by good laboratory practice. Contaminated reagents in laboratories have led to false positive test results.[42] A proportion of samples within the study should therefore be tested for cross contamination, and this proportion should be stated.

Measures of precision (repeatability and reproducibility) might 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 (eg, time, operators, calibrators, and reagent lots).[43] The lot-to-lot variability of tests should be stated.

Postanalytical characteristics—decision limits

Decision limits need to be defined for positive, negative, and indeterminate results. Preferably, these cut-off 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-off points derived from the data collected within the study can bias estimates of test performance.[44 45] If no prior data exist to determine cut-off points, or when the cut-off point was established in people with symptoms but the test is intended to be used in non-symptomatic individuals or individuals with mild symptoms, 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, considerable attention has been given to the number of amplification cycles used and the cycle threshold (C_T) to determine if a test is positive, negative, or indeterminate. Although a strong relation exists between C_T and viral load, choosing the C_T is not easily generalisable between tests, kits, testing platforms, and laboratories. C_T values can 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-off points 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 can be between different forms of testing, 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 can 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 first scenario, both index tests are best performed in the same individuals, using a direct comparison, rather than as 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 could be to perform a descriptive analysis of all included index tests or to determine if a new test has higher sensitivity and equivalent specificity, or faster turnaround time and equivalent diagnostic accuracy. Although one characteristic might be specified as the primary outcome (eg, 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 individuals who have the target condition from those who do not; for example, those who have or have had the infection from those who do not or 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 (or tests), the comparator test (or tests), and the reference standard test need to be conducted masked to the results of the other test (or tests).

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, which does not apply to the clinical population of interest.[9-15] 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.[46]

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

SARS-CoV-2 infection includes individuals who do not have symptoms, those who are presymptomatic, and those who have symptoms. WHO has published definitions of suspected, probable, and confirmed covid-19 based on clinical, epidemiological, and laboratory criteria, with recommended associated testing.[47 48] According to this advice, a person with confirmed covid-19 is defined as having laboratory confirmation of SARS-CoV-2 infection, irrespective of clinical signs and symptoms. This definition can be confusing, because 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 in itself.

WHO defines an individual with probable covid-19 as having symptoms indicative of the disease (fever, cough, general weakness or fatigue, headache, myalgia, sore throat, coryza, dyspnoea, anorexia, nausea or vomiting, diarrhoea, altered mental status); has an epidemiological risk of exposure; and is a contact of a person with probable or confirmed covid-19, has chest imaging findings suggestive of covid-19, has a loss of taste or smell, or death has occurred that is not otherwise explained in an adult with respiratory distress preceding death and was a contact of an individual with probable or confirmed covid-19 or epidemiologically linked to a cluster with at least one person with confirmed covid-19. These 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 having probable covid-19, but not be infected with SARS-CoV-2. On the other hand, some individuals will have had exposure, have had symptoms and investigations such as imaging that indicate covid-19, but have tested (either by RT-PCR or antibody) negative. These individuals are not classified as having definite covid-19. If the WHO classification is used as a reference standard, a sensitivity analysis of the test's clinical performance using a reference standard including probable disease should be presented.

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 two weeks before testing. After this period, viral load decreases in many individuals reducing the sensitivity of the RT-PCR. Although the specificity of viral nucleic acid testing is thought to be 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 might be due to cross contamination with other samples or clerical error in reporting results. Repeat testing could 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 after exposure, might be negative because the individual no longer has the virus. Repeat testing in individuals with confirmed covid-19 shows that false negative results occur, particularly in the first few days after exposure or late in the course of infection.[34-36 49 50] Poor sampling technique, samples from the wrong anatomical site, and incorrect transport of specimens can also contribute to false negative results. One negative viral nucleic acid test is inadequate to rule out SARS-CoV-2 infection.

Performance of viral nucleic acid testing as a reference standard could be improved by ensuring appropriate collection, repeat testing for those who initially test negative within an appropriate time window (eg, within five days after symptom onset or on the fourth day after exposure if exposure date is known), or by samples from multiple sites or with multiple genetic targets.[51 52] Serology could be used if exposure is thought to have occurred more than 14 days previously. However, serology also has a high false negative rate, and might 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.[53 54] Repeat testing and combinations of tests, however, adds 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 with a discordant result between the index test and the reference standard.[55]

For asymptomatic infection, clinical reference standards are not possible because there are no clinical symptoms and because 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 or 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 individuals with probable disease. 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.[13] 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 underestimation of the diagnostic sensitivity of an index test (if individuals are defined by the reference standard as having disease are actually true negatives) or an overestimation 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. Such evidence can be shown through results of a previous 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 presymptomatic individuals. Therefore, if the test is intended for seroprevalence surveys, the best study design would involve a large number of randomly selected individuals 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 category for individuals with probable disease category might be useful.

For studies where the target condition is infectiousness

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—that is, are still viable virus particles as opposed to inactive or remnants of virus particles. The ability to use a rapid test that determines whether 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 because they appear to have suboptimal sensitivity for detecting infectiousness. Early in the course of infection, which we expect to be the most infectious stage, samples from RT-PCR positive individuals with the virus might not grow virus on cell culture.[56] While samples that return a positive RT-PCR result at a higher C_T could indicate viral remnants at a point when the patient is no longer infectious, they might also indicate an early point in the course of the infection. Using a lower C_T for determining infectiousness will reduce the sensitivity of the test to detect all infectious individuals. Similarly, the assumption that only those people with high viral load are infectious will miss individuals who have lower viral loads but who are still capable of passing on the infection.[16]

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

Diagnosis of SARS-CoV-2 clearance (that is, 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.[12] Time for clearance from gastrointestinal tract varies greatly by individual. It is not known whether 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.[8] Researchers should include the STARD flow diagram to report the number of individuals included in the study, the number of individuals excluded from the study before 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 (eg, who did not receive the reference standard, or had indeterminate or outlier results; fig 3). The diagram might need to be adapted for 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 one test), then the sample size is again the number of individuals included in the study.

Although researchers should evaluate sensitivity and specificity in the same population to estimate clinical test performance, preliminary studies might 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, a cross tabulation of the index test and the reference standard results is helpful. Use of the same reference standard for all index tests minimises the risk of verification bias. Any missing data or indeterminate results for either the index test or reference standard should be reported according to the final disease status (if known) and not excluded from the results.

Reports can include the results of analytical performance (eg, analytical sensitivity, analytical specificity, imprecision), 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.[57 58]

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 might include a RT-PCR test 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 might be defined as time since RT-PCR evaluation, or better, the time since exposure to a known person with confirmed disease 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.

Subgroup analyses

Subgroup 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. Subgroup 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 or presymptomatic or those who have symptoms are particularly helpful.

Comparative analyses

As described above, two index tests should ideally 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 might be reported in the form of tables showing concordant and discordant results. Further information on the people with discordant results could help to evaluate which test is more accurate using agreement with observations that might be considered as so-called fair umpires but are not a reference standard.[59] Such fair umpires could include information on prior exposure risk, concurrent tests (apart from index or comparator test under evaluation—eg, inflammatory markers, chest imaging), response to treatment, and clinical outcomes on follow-up.

Predictive values

Clinicians and public health experts require not only the sensitivity and specificity of the test but also an understanding of the positive and negative predictive values of the test. In presenting the results of the study, estimates of these predictive values using several clinically relevant values of prevalence is helpful. We also recommend a graphical display of how the test characteristics will perform in slightly different prevalence settings and use of natural frequencies (eg, the number of people affected in a population of 10 000 people), as shown in figure 1. The FDA website provides 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.[60]

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 might need to be refined on the basis of understanding gained from the evaluation of the test.

If a study is done in a reference laboratory with highly experienced staff, 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.[61] 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 understand the clinical performance of tests to date, including having to withdraw tests from the market that have been shown to have poor test accuracy.[62 63] 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[64] or using inaccurate rapid testing to screen asymptomatic individuals and falsely reassuring individuals who are infectious.[16] 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.[65 66] 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. The use of convenience samples and retrospectively collected samples has been a particular problem for the evaluation of antibody tests.[9] Submissions for emergency use authorisation 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 organisations, regulatory authorities, and scientific journal editorial boards could assist by harmonising their requirements for test evaluations and developing study templates that can be used across studies and that encourage standardised data collection and reporting and rigorous study design.

Author affiliations

[1]Centre for Longitudinal and Life Course Research, School of Public Health, University of Queensland, Herston, QLD 4006, Australia

[2]School of Public Health, University of Sydney, NSW, Australia

[3]Department of Epidemiology and Data Science, Amsterdam University Medical Centres, University of Amsterdam, Amsterdam, Netherlands

[4]Test Evaluation Research Group, Institute of Applied Health Research, University of Birmingham, Birmingham, UK

[5]NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust and University of Birmingham, Birmingham, UK

[6]School of Medicine, Sydney, University of Notre Dame, Darlinghurst, NSW, Australia

[7]Centre for Medical Imaging, University College, London, UK

[8]Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht University, Utrecht, Netherlands

[9]Institute of Infection, Veterinary, and Ecological Sciences, University of Liverpool, Liverpool, UK

[10]Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway

[11]Norwegian Quality Improvement of Laboratory Examinations, Haraldsplass Deaconess Hospital, Bergen, Norway

[12]CALIPER Program, Paediatric Laboratory Medicine, The Hospital for Sick Children, Toronto, ON, Canada

[13]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada

[14]New South Wales Health Pathology, Department of Chemical Pathology, Prince of Wales Hospital, Sydney, NSW, Australia

[15]School of Medical Sciences, University of New South Wales, Sydney, NSW, Australia

We thank Tze Ping Loh, Mary Kathryn Bohn, and Shannon Steele for their helpful assistance and comments on the manuscript, and Ian Doust for providing figure 2.

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. 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 corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: No specific funding was provided for this work. 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 (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. 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 views expressed are those of the authors and not necessarily those of the NHS, NIHR, or Department of Health and Social Care.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Provenance and peer review: Not commissioned; externally peer reviewed.

Patient and public involvement: 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.

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;395:565-74. doi:10.1016/S0140-6736(20)30251-8

2 Zhang Y-Z. Novel 2019 coronavirus genome. Virological 2020. https://virological.org/t/novel-2019-coronavirus-genome/319

3 Foundation for Innovative New Diagnostics. Homepage. 2021https://www.finddx.org/covid-19/

4 European Commission. JRC Covid-19 In Vitro Diagnostic Devices and Test Methods. https://covid-19-diagnostics.jrc.ec.europa.eu/

5 Norwegian Institute of Public Health. NIPH systematic and living map on COVID-19 evidence. https://www.nornesk.no/forskningskart/NIPH_diagnosisMap.html

6 Tromberg BJ, Schwetz TA, Pérez-Stable EJ, et al. Rapid Scaling Up of Covid-19 Diagnostic Testing in the United States - The NIH RADx Initiative. *N Engl J Med* 2020;383:1071-7. doi:10.1056/NEJMsr2022263

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.

8 Bossuyt PM, Reitsma JB, Bruns DE, et al, STARD Group. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 2015;351:h5527. doi:10.1136/bmj.h5527

9 Deeks JJ, Dinnes J, Takwoingi Y, et al, Cochrane COVID-19 Diagnostic Test Accuracy Group. Antibody tests for identification of current and past infection with SARS-CoV-2. *Cochrane Database Syst Rev* 2020;6:CD013652. doi:10.1002/14651858.CD013652

10 Lisboa Bastos M, Tavaziva G, Abidi SK, et al. Diagnostic accuracy of serological tests for covid-19: systematic review and meta-analysis. *BMJ* 2020;370:m2516. doi:10.1136/bmj.m2516

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

12 Mallett S, Allen AJ, Graziadio S, et al. 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;18:346. doi:10.1186/s12916-020-01810-8

13 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 [Preprint]. https://www.medrxiv.org/content/10.1101/2020.04.16.20066787v2.article-info

14 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;58:102916. doi:10.1016/j.ebiom.2020.102916

15 Dinnes J, Deeks JJ, Adriano A, et al, 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;8:CD013705. doi:10.1002/14651858.CD013705

16 Deeks JJ, Raffle AE. Lateral flow tests cannot rule out SARS-CoV-2 infection. *BMJ* 2020;371:m4787. doi:10.1136/bmj.m4787

17 Whiting PF, Rutjes AW, Westwood ME, et al, QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011;155:529-36. doi:10.7326/0003-4819-155-8-201110180-00009

18 US Food and Drug Administration. In Vitro Diagnostics EUAs. 2021. https://www.fda.gov/medical-devices/coronavirus-disease-2019-covid-19-emergency-use-authorizations-medical-devices/vitro-diagnostics-euas

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. 2021. 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

20 World Health Organization. Target product profiles for priority diagnostics to support response to the COVID-19 pandemic version 1.0. 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

22 EUR-Lex. 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

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

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

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

26 Infectious Diseases Society of America. Guidelines on the Diagnosis of COVID-19. 2020. https://www.idsociety.org/COVID19guidelines/dx

27 Stephens DS, McElrath MJ. COVID-19 and the Path to Immunity. *JAMA* 2020;324:1279-81. doi:10.1001/jama.2020.16656

28 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 [Preprint]. https://www.medrxiv.org/content/10.1101/2020.07.14.20151126v1

29 Sewell HF, Agius RM, Stewart M, Kendrick D. Cellular immune responses to covid-19. *BMJ* 2020;370:m3018. doi:10.1136/bmj.m3018

30 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;5:e452-9. doi:10.1016/S2468-2667(20)30157-2

31 Woloshin S, Patel N, Kesselheim AS. False Negative Tests for SARS-CoV-2 Infection - Challenges and Implications. *N Engl J Med* 2020;383:e38. doi:10.1056/NEJMp2015897

32 Lord SJ, St John A, Bossuyt PM, et al, 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;56:527-35. doi:10.1177/0004563219842265

33 Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. *BMJ* 2006;332:1089-92. doi:10.1136/bmj.332.7549.1089

34 He X, Lau EHY, Wu P, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat Med* 2020;26:672-5. doi:10.1038/s41591-020-0869-5

35 Wölfel R, Corman VM, Guggemos W, et al. Virological assessment of hospitalized patients with COVID-2019. *Nature* 2020;581:465-9. doi:10.1038/s41586-020-2196-x

36 Walsh KA, Jordan K, Clyne B, et al. SARS-CoV-2 detection, viral load and infectivity over the course of an infection. *J Infect* 2020;81:357-71. doi:10.1016/j.jinf.2020.06.067

37 Lee S, Kim T, Lee E, et al. 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;e203862. doi:10.1001/jamainternmed.2020.3862

38 Gudbjartsson DF, Norddahl GL, Melsted P, et al. Humoral Immune Response to SARS-CoV-2 in Iceland. *N Engl J Med* 2020;383:1724-34. doi:10.1056/NEJMoa2026116

39 Service RF. Fast, cheap tests could enable safer reopening. *Science* 2020;369:608-9. doi:10.1126/science.369.6504.608

40 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;58:e00796-20. doi:10.1128/JCM.00796-20

41 Theel E, Filkins L, Palavecino E, et al. Verification procedure for commercial serologic tests with Emergency Use Authorization for detection of antibodies to SARS-CoV-2. American Society for Microbiology. 2020. https://asm.org/Protocols/Verify-Emergency-Use-Authorization-EUA-SARS-CoV-2.

42 Willman D. Contamination at CDC Laboratory Delayed Rollout of Coronavirus Tests. *Washington Post* 2020.https://www.washingtonpost.com/investigations/contamination-at-cdc-lab-delayed-rollout-of-coronavirus-tests/2020/04/18/fd7d3824-7139-11ea-aa80-c2470c6b2034_story.html.

43 Clinical and Laboratory Standards Institute. Harmonized Terminology Database. 2020. https://htd.clsi.org/listterms.asp?searchd.

44 Ewald B. Post hoc choice of cut points introduced bias to diagnostic research. *J Clin Epidemiol* 2006;59:798-801. doi:10.1016/j.jclinepi.2005.11.025

45 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. *medRxiv* 2020 [Preprint]. https://www.medrxiv.org/content/10.1101/2020.07.21.20159038v1

46 Bossuyt PM. Testing COVID-19 tests faces methodological challenges. *J Clin Epidemiol* 2020. doi:10.1016/j.jclinepi.2020.06.037

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

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

49 Long C, Xu H, Shen Q, et al. Diagnosis of the Coronavirus disease (COVID-19): rRT-PCR or CT?*Eur J Radiol* 2020;126:108961. doi:10.1016/j.ejrad.2020.108961

50 Kucirka LM, Lauer SA, Laeyendecker O, Boon D, Lessler J. Variation in False-Negative Rate of Reverse Transcriptase Polymerase Chain Reaction-Based SARS-CoV-2 Tests by Time Since Exposure. *Ann Intern Med* 2020;173:262-7. doi:10.7326/M20-1495

51 Pan Y, Zhang D, Yang P, Poon LLM, Wang Q. Viral load of SARS-CoV-2 in clinical samples. *Lancet Infect Dis* 2020;20:411-2. doi:10.1016/S1473-3099(20)30113-4

52 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. doi:10.1016/j.cmi.2020.05.013

53 Zou MY, Wu GQ. The effect of antigen cross-reaction on testing of SARS-CoV-2 specific antibodies in serum. *Chinese J Clin Lab Sci* 2020;3:161-3.

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

55 Hadgu A. Discrepant analysis: a biased and an unscientific method for estimating test sensitivity and specificity. *J Clin Epidemiol* 1999;52:1231-7. doi:10.1016/S0895-4356(99)00101-8

56 Bullard J, Dust K, Funk D, et al. Predicting infectious SARS-CoV-2 from diagnostic samples. *Clin Infect Dis* 2020;ciaa638.

57 Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Stat Med* 1998;17:857-72. doi:10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E

58 Wilson EB. Probable Inference, the Law of Succession, and Statistical Inference. *J Am Stat Assoc* 1927;22:209-12. doi:10.1080/01621459.1927.10502953

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

60 Food and Drug Administration. EUA Authorized Serology Test Performance. 2021. https://www.fda.gov/medical-devices/emergency-situations-medical-devices/eua-authorized-serology-test-performance

61 Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* 2016;6:e012799. doi:10.1136/bmjopen-2016-012799

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

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

64 Calafiore S. GPs face $20k fines for using serology tests to diagnose coronavirus. 2020. 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

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

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