NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE

Evidence standards framework for SARS-CoV-2 and anti-SARS-CoV-2 antibody diagnostic tests

The framework

This framework is a 3-stage approach to collecting the best possible data and evidence in the short and long term, while tests are being quickly developed and validated during the COVID-19 pandemic.

The framework aligns with the UK government's 5-pillar testing strategy by including viral detection tests for current infection (pillars 1 and 2) and antibody tests for previous infection (pillars 3 and 4).

We've tried to ensure that it builds on other key national policy documents and requirements, such as Medicines & Healthcare products Regulatory Agency (MHRA) guidance on COVID-19, including their target product profiles and technical specifications.

The framework also draws upon internationally recognised tools and reporting standards for diagnostic test accuracy studies such as QUADAS-2 and STARD.

The framework is not intended to be exhaustive, but provides some important issues to consider.

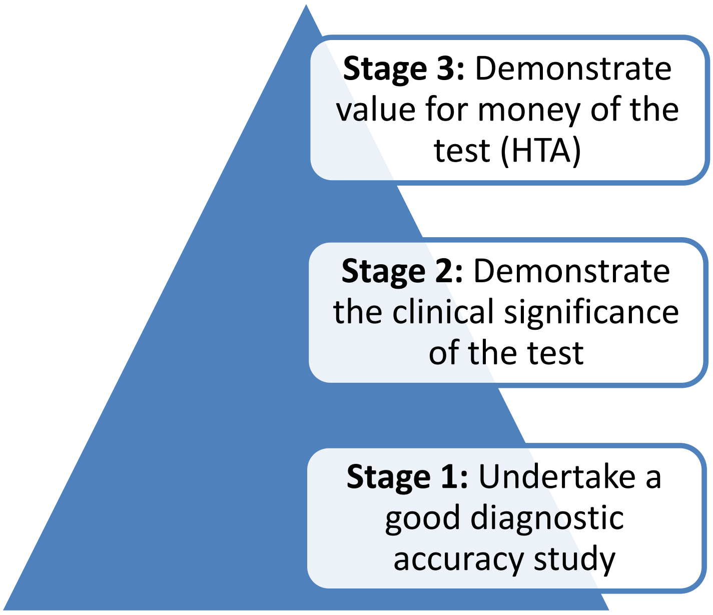
Who the framework is for

It is for anyone working on testing for COVID-19, in particular:

- diagnostic test manufacturers
- laboratories developing tests
- clinicians assessing tests
- clinical researchers collecting data on tests
- research funders and people who advise researchers on study design
- purchasers and other decision makers.

The 3 stages

The 3-stage approach is summarised in the diagram and explained in the text that follows.


Stage 3: Demonstrate value for money of the test (HTA)
Stage 2: Demonstrate the clinical significance of the test
Stage 1: Undertake a good diagnostic accuracy study

The framework assumes that the tests' analytical performance is already established, and that developers are complying with existing quality systems for manufacturers (ISO 13485) and laboratories (ISO 15198 or 17025).

Stage 1: doing a good diagnostic test accuracy study

Sensitivity and specificity

Diagnostic test performance is assessed based on clinical:

- sensitivity (the proportion of people with a condition who test positive)
- specificity (the proportion of people without a condition who test negative).

The MHRA's target product profiles require clinical sensitivity and specificity estimates for tests.

Quality assessment

The QUADAS-2 tool gives criteria for quality assessing a diagnostic accuracy study. The tool helps identify where bias could be introduced in a study and how to make sure the results are relevant to how the test is intended to be used in the UK.

Reporting

Report studies clearly and transparently, with enough information on how they were carried out. The STARD reporting guideline has a checklist of what should be included in reports of diagnostic accuracy studies.

Test information

Clearly explain the intended use of the test (in a diagnostic test accuracy study, or in literature accompanying a test). For example:

- which population and sample type the test is for
- the setting(s) where it will be used
- what type(s) of healthcare professional administers the test
- when the test should be used (for example, a certain number of days after symptoms)
- what happens after the test is done (for example posted, shipped to lab)
- the purpose of the test
- what a positive and negative test result means.

Study design

The standard diagnostic accuracy study is usually a cohort design: all participants get both the test being assessed (the index test) and the reference standard test. But this may not be possible during the COVID-19 pandemic. For example, if the test is only available in small quantities, or the condition's prevalence is low, making samples sizes too large.

Holtman et al. (2019) sets out alternative study designs for diagnostic accuracy studies that can be used in low prevalence populations. These can bias test accuracy values, so consult a statistician about how to minimise bias in these studies.

The study design may need to be dictated by the reference standard. For example, in people negative for anti-SARS-CoV-antibodies the best reference standard is a historic sample from before the COVID-19 pandemic, so a study using known negative and positive cases may be the only feasible design. Using the QUADAS-2 tool will help to reduce the bias that this design risks introducing, for example, in how the index test is done and interpreted. Matching the positive and negative cases to the target population as closely as possible also helps improve the generalisability of the results (see 'study population'). Some aspects of the QUADAS-2 tool may not be relevant for this study design, for example, the reference standard used will differ between positive and negative cases.

Prognostic accuracy studies may also be important. See other outcomes in this document.

Study population

Inclusion and exclusion criteria should be clearly specified. Inappropriate exclusion criteria that excludes people who would be offered the test when used in its intended population in the UK should be avoided. This is a risk if people are excluded based on characteristics, such as comorbidities, that mean they are likely to be difficult to diagnose.

Participants enrolled should match the population(s) that the test is for (in the UK) as closely as possible. For example, in terms of:

- condition severity, for example the extent of symptoms of possible COVID-19
- presence of other conditions
- the amount of any previous testing
- demographic features.

You can enrol all participants meeting the inclusion criteria who consent to testing after the start of the study, or a random sample of them.

A two-gate case-control trial design assesses the test in people known to have COVID-19 (or an immune response) and people known not to have COVID-19 (or an immune response). This design may overestimate diagnostic accuracy because people with a milder condition may be excluded from the known COVID-19 sample population. If this design is used, the positive case samples should match the range of disease in the population of interest as closely as possible.

For example, for tests detecting SARS-CoV-2 RNA, positive cases should include people with a likely range of viral load from low to high, if this is expected in the setting the test will be used in.

The performance of the test in subgroups, such as people with comorbidities, should be considered.

The index test

The test that is assessed in the study is called the index test.

The index test should be interpreted without knowledge of the reference standard result, unless the result of the test does not require any judgement (subjectivity) by the interpreter.

If a threshold value (or cut-off value; a value the test result has to be over or under to make a positive diagnosis) is used to determine a positive result, this should be specified before starting the study. It should be the value recommended for using the test in the UK.

The index test should be done and interpreted as closely as possible to how the test will be used in the UK. For example, by:

- using the version of the test available in the UK

- using the recommended protocol
- making sure the person doing and interpreting the test has similar experience to the intended UK user, for example a trained healthcare professional or a member of the public.

The reference standard

The choice of reference standard can have a large impact on accuracy estimates. The test accuracy of an index test is calculated using an assumption that the reference standard is 100% accurate (a ‘gold standard’). If this is not true, it will affect estimates of accuracy.

If a reference standard wrongly indicates someone does not have a condition, and the index test correctly indicates they do, this is a false positive index test result.

If both reference and index tests wrongly indicate someone does not have a condition, the index test is wrongly considered to give a true negative result.

The reference standard should, therefore, be the best available measure of the condition the index test is for.

If no one acceptable reference standard is available, consider a composite reference standard with results from several tests used to determine a positive result based on agreed criteria. Alternatively, consider using latent class models, consulting with a statistician and clearly reporting the statistical approach. The STARD-BCLM is a checklist for reporting diagnostic accuracy studies that use Bayesian latent class models.

The reference standard test should be interpreted without knowing the result of the index test, unless the result of the test does not require any judgement (subjectivity) by the interpreter.

All diagnostic accuracy studies of COVID-19 tests should explicitly define what reference standard was used to identify true and false positive and negative cases. Ideally, consult with clinicians to help identify the most appropriate reference standard.

How people are tested with the index and reference tests (flow and timing)

The tests should be done at the same time unless there is a reason to do them at different times. That is, there should be no difference in how, or how long, the sample was stored between index and reference tests.

If by design the reference standard or index tests are done at different times, explain the length of the time interval. For example, the presence of antibodies may need to be assessed several days or weeks after a confirmed SARS-CoV-2 infection (the reference standard).

Clear reasons should be provided for any difference in the number of people enrolled in the study and the number the test(s) were done on. If people were lost to follow up before testing could be done (either reference standard, index test or both), reasons for this should be provided.

If not everyone who had a reference standard had an index test (or vice versa) through study design (see study design in this document), the methodology for selecting who had tests should be clearly stated. The number of indeterminate test results should also be reported, as should the number of test failures (when no result was produced by the test).

Reporting test results

Individual numbers of people with true positive, false positive, true negative and false negative test results should be reported, not just the derived accuracy estimates from these figures, for example sensitivity and specificity.

Results can be reported in a 2x2 table showing the numbers of individual participants whose index and reference tests agreed and disagreed (for positive and negative results), for example:

Test	Index test: positive	Index test: negative
Reference standard: positive	True positive	False negative
Reference standard: negative	False positive	True negative

If possible, individual patient data should be made available (with information governance procedures in place), linked to the results of the index and reference tests. For example:

- whether a person had symptoms
- time since symptoms or confirmed diagnosis of COVID-19 (for antibody tests)
- sex
- age
- comorbidities
- the length of time between index and reference test.

Other outcomes

Test accuracy studies can also collect data on other potentially valuable outcomes. These data can be useful to determine what impact the test result can have on clinical care and how easy it is to implement, for example:

- the time from taking a sample to getting a test result
- how easy it is to do and interpret the test
- the test failure rate
- the impact of the test result on clinical decision making, for example changes to care because of it.

Prognostic accuracy studies can also be useful, for example a study to explore if viral load correlates with COVID-19 severity or prognosis. The PROBAST tool helps assess the risk of bias and applicability of diagnostic and prognostic prediction model studies.

Stage 2: demonstrating clinical significance

Clinical significance of test results

Accuracy measurements assess the test against a reference standard that is the best measurement available. However, the accuracy of a test does not show if:

- people who are tested have better outcomes than those who are not tested or who have an alternative test
- rates of transmission and infection reduce as a result of testing compared with no testing or an alternative test.

The purpose of stage 2 in this framework is to gather data on any intermediate measures or clinical outcomes. These relate to changes in care decisions, or changes in treatments or actions (such as returning to work or self-isolating) as a result of the test. During this stage, longer-term follow-up outcomes from people who were tested that are relevant to patients or to wider society should be reported.

Ideally data will be collected through a randomised controlled trial in which:

- one group of patients is treated or modifies its behaviour based on the results from the new test
- another group of patients is treated or modifies its behaviour based on the results of a different test, clinical judgement, or symptoms.

Alternatively, studies could look at:

- the impact on treatment decisions or behaviour after test results
- clinical outcomes from different treatments or transmission rates from different models of behaviour.

Outcomes from these studies could then be linked using modelling during stage 3 of the framework.

Here are some examples of potential studies and outcomes to demonstrate clinical significance.

Example 1: for viral detection tests to diagnose people presenting to hospital with suspected COVID-19, decision impact studies will help understand what management decisions are made based on the test result, and what care and assessment people get after their diagnosis. Clinical studies of the different treatments used for people with COVID-19 could report outcomes such as:

- length of hospital stay
- severity of symptoms
- mortality
- resource use.

This will help to show the benefits of correctly diagnosing COVID-19, and the impact of missing a diagnosis of the condition.

Example 2: for antibody tests, studies may include longer-term follow up of people with positive and negative test results to monitor outcomes such as:

- behaviour modification, for example decisions about social interactions
- rates of COVID-19 infection or reinfection
- quality of life
- usability of test kits
- failure rates in practice.

Example 3: for viral detection tests used in a population of key workers who are self-isolating because of symptoms, studies should report on behaviour modification, such as whether people follow isolation rules following a positive test result and if they return to work after a negative test result.

Stage 3: developing the value proposition

Health technology assessment of tests

Stage 3 looks at the longer-term implications of test results on costs and health-related or societal outcomes to estimate whether the tests provide value for money.

This could be done by combining outputs from stage 1 and 2 using linked evidence-modelling studies. This would generate the evidence needed for a full health technology assessment. If the accuracy estimates from stage 1 are uncertain, this should be explored in the modelling.

Outcome studies that follow patients from testing through to end health outcomes can also be used in a health technology assessment.

False positives and negatives, and the potential implications of acting (incorrectly) on a false test result, are particularly useful in this context. For example, in the case of viral detection testing, a false positive in a person with mild or no symptoms may be stressful and lead to unnecessary behaviour changes, such as staying away from work.

Conversely, a false negative might not impact the person tested much if symptoms are mild and no treatment is needed. But it has a societal impact because they may go on to infect several other people.

For antibody testing, a false positive result could give false reassurance that the person will not get infected with COVID-19.

Modelling studies should show what impact this has on:

- patient-relevant outcomes
- societal outcomes (such as employment and productivity)
- resource use
- onward transmission rates.

Societal impacts are not typically considered by NICE in its reference case but can be relevant for other groups carrying out health technology assessments.

Why NICE developed an evidence standards framework for COVID-19 diagnostic tests

Testing is central to the UK government's strategy to tackle COVID-19. The government's 5-pillar testing strategy includes viral detection tests to find out if someone has the virus, and antibody tests to find out if they've had it before.

The 5-pillar strategy says that all tests that are used need to be validated, reliable and accurate. Industry, academia and Public Health England have been working rapidly to develop and validate tests. NICE is supporting them with this framework.

How the framework was developed

This framework was developed by NICE's Diagnostics Assessment Programme and revised in response to comments from selected stakeholders and experts. In developing the framework, we have tried to balance the need for rapid data collection and evidence generation with the need for accurate and robust data and evidence.

Updates to the framework

The framework will be reviewed during the pandemic and may require further rapid update in response to changes in COVID-19 diagnostic testing, for example as we learn more about the methods for evaluating antibody tests and the significance of immunity becomes better understood.

Disclaimer

The advice in this evidence standards framework is based on the scientific and methodological knowledge publicly available at the time of writing and cannot account for future changes and developments in scientific knowledge, regulatory requirements, or any referenced material from external sources.

The content of this document may be subject to further revisions as more scientific and methodological knowledge becomes available. NICE cannot accept responsibility or liability for the use of its content in third-party outputs, services or related healthcare settings.