Lessons to be learned from test evaluations during the Covid-19 pandemic

Deeks, Jonathan J; Ashby, Deborah; Takwoingi, Yemisi; Perera, Rafael; Evans, Stephen JW; Bird, Sheila M

DOI:
10.1093/jrsssa/qnae053

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Deeks, JJ, Ashby, D, Takwoingi, Y, Perera, R, Evans, SJW & Bird, SM 2024, 'Lessons to be learned from test evaluations during the Covid-19 pandemic: RSS Working Group's Report on Diagnostic Tests', *Journal of the Royal Statistical Society Series A (Statistics in Society)*, pp. 1-51. https://doi.org/10.1093/jrsssa/qnae053

Journal of the Royal Statistical Society Series A: Statistics in Society, 2024, **00**, 1–51
https://doi.org/10.1093/jrsssa/qnae053

Journal of the Statistics Society
Series A
Statistics in Society

Working Party Report

Lessons to be learned from test evaluations during the COVID-19 pandemic: RSS Working Group's Report on Diagnostic Tests

Jonathan J. Deeks[1], Deborah Ashby[2], Yemisi Takwoingi[1], Rafael Perera[3], Stephen J.W. Evans[4] and Sheila M. Bird[5,6]

[1]Institute of Applied Health Research, University of Birmingham, Birmingham, UK
[2]School of Public Health, Imperial College London, London, UK
[3]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK
[4]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK
[5]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK
[6]University of Edinburgh College of Medicine and Veterinary Medicine, Edinburgh, UK

Address for correspondence: Jonathan J. Deeks, Institute of Applied Health Research, Public Health Building, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK. Email: j.deeks@bham.ac.uk

Abstract

The coronavirus disease (Covid-19) pandemic raised challenges for everyday life. Development of new diagnostic tests was necessary, but under such enormous pressure risking inadequate evaluation. Against a background of concern about standards applied to the evaluation of in vitro diagnostic tests (IVDs), clear statistical thinking was needed on the principles of diagnostic testing in general, and their application in a pandemic. Therefore, in July 2020, the Royal Statistical Society convened a Working Group of six biostatisticians to review the statistical evidence needed to ensure the performance of new tests, especially IVDs for infectious diseases—for regulators, decision-makers, and the public. The Working Group's review was undertaken when the Covid-19 pandemic shone an unforgiving light on current processes for evaluating and regulating IVDs for infectious diseases. The report's findings apply more broadly than to the pandemic and IVDs, to diagnostic test evaluations in general. A section of the report focussed on lessons learned during the pandemic and aimed to contribute to the UK Covid-19 Inquiry's examination of the response to, and impact of, the Covid-19 pandemic to learn lessons for the future. The review made 22 recommendations on what matters for study design, transparency, and regulation.

Keywords: covid-19, in vitro diagnostic test, scientific integrity, statistics, test evaluation

1. Introduction

In January 2020, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was identified causing coronavirus disease (Covid-19), and began its rapid global spread, with the first UK lockdown commencing on 23rd March 2020. Science and scientists moved at great pace to combat the pandemic infection. Statistics played a key role in scientific and medical developments.

Key areas of statistical input included modelling to predict the spread under various assumptions, surveillance studies regularly to monitor changes in the prevalence of current or previous infection (nationally and regionally), development of strategies for testing, tracing and isolation, trials of treatments in a variety of settings, and of vaccines. As each of these has multiple statistical issues, the Royal Statistical Society's (RSS) Council set up a Covid-19 Task Force (Royal Statistical

Received: February 1, 2024. **Accepted:** May 2, 2024

Society, 2020) with a key aim of ensuring that the RSS could contribute its collective expertise on statistical issues during the Covid-19 pandemic to the UK's national and devolved governments and public bodies.

Because the pandemic developed so quickly, there was enormous pressure for rapidly available solutions, but at the risk of inadequate evaluation. The RSS was particularly concerned that many new diagnostic tests for SARS-CoV-2 were coming to market for use both in clinical practice and for surveillance without adequate provision for statistical evaluation of their analytical and clinical performance. Against a background of concern about standards applied to the evaluation of in vitro diagnostics (IVDs), there was a need for clear statistical thinking on the principles of diagnostic testing in general, and their application in a pandemic. Therefore, in July 2020, the RSS convened a Working Group of six biostatisticians experienced in regulation and test evaluation (members are listed in the Supplementary Material). The Terms of Reference of the Working Group were to review the statistical evidence needed to assure the performance of new tests, for patients, decision-makers and regulators, with particular reference to IVDs for infectious diseases. Considerations included:

- Statistical issues specific to the diagnosis and surveillance of infectious diseases, including new emerging infectious diseases;
- Key characteristics to be evaluated when assuring the performance of an IVD test for an infectious disease;
- Design aspects of studies that are necessary to provide estimates of these key characteristics;
- Statistical principles to be followed by decision makers (including regulators) when assessing the adequacy of performance of a test for its intended role in the protection of public health;
- Information that needs to be in the public domain to provide confidence in the performance of tests.

The Covid-19 pandemic provided a microcosmic insight into the inadequate state of current processes for evaluating and regulating medical tests. Whilst directly motivated by the pandemic, the findings apply more broadly to IVDs and, in part, to all diagnostics.

The report contains seven numbered sections. Section 2 provides a background to infectious diseases and key terminology, and Section 3 outlines key concepts in diagnostic testing, with a more detailed exposition of statistical estimands in Section 4. Section 5 addressed considerations of good study design for the evaluation of diagnostic tests. Section 6 addressed the Covid-19 pandemic specifically, and the lessons to be drawn through use of 15 examples. Section 7 deals with the implications for the regulation of diagnostic tests, and Section 8 discusses the information that should be in the public domain. Supplementary Material includes a Glossary of technical terms and an Appendix of examples of major infectious diseases.

The Working Group identified 22 recommendations, organized under headings of Study Design (Box 1), Regulation (Box 2), and Transparency (Box 3). The original version of the Working Group's report was published on the RSS website and disseminated to the public in June 2021. This updated version of the report includes additional information on *Assessing Test Strategies and their Impact* (Section 6.10). The RSS Working Group report has been submitted for the UK Covid-19 Inquiry's examination of the UK's response to the Covid-19 pandemic, to learn lessons for the future.

2. Understanding infectious diseases

Emerging and existing infectious diseases are a threat to global health. Increased virulence, incidence, geographic distribution or the development of drug resistance intensifies the challenge of existing infectious diseases in public health systems around the world.

Infectious organisms (pathogens) include viruses (e.g. SARS-CoV-2), bacteria (e.g. *Mycobacterium tuberculosis*), parasites (e.g. *Plasmodium* species), and fungi (e.g. *Candida* species).

Pathogens have different consequences in terms of their morbidity and mortality, social and economic impacts. Each year, several outbreaks of infectious diseases are reported in different parts of the world (World Health Organization, 2024). Notable recent outbreaks include SARS

Box 1: Recommendations—Study-design matters.

1. Robust studies of analytical performance provide necessary but insufficient evidence to implement in vitro diagnostics.
2. Field or clinical evaluation studies are needed to evaluate the performance of an in vitro diagnostic for each intended use.
3. Definition of each intended use requires specification of: (a) the people, place and purpose of testing; (b) the target condition that testing aims to detect; (c) the test's specimen-type and how the specimen is taken, stored and transported and by whom; and (d) details of the individuals, training and facilities where testing is done.
4. Undertaking well designed, adequately powered, and correctly analysed studies of the clinical performance of an in vitro diagnostic is important for each intended use of the test. Study completion may be easier and faster in pandemics because of the rapid accrual of cases.
5. Consideration of sensitivity (% of infected persons who are correctly detected by the test) and specificity (% of uninfected persons correctly labelled by the test as uninfected) for each intended use should be *de rigueur*, not exceptional.
6. It is important to know the likely prevalence of the condition in the target population to be able to ascertain the probability that a positive test result is correct (the positive predictive value) and that a negative test result is correct (the negative predictive value).
7. To quantify sampling uncertainty, estimates of prevalence and test performance must be presented with confidence intervals (or other appropriate measures).
8. Direct comparison of alternative in vitro diagnostics and test strategies should be given high consideration to provide evidence that directly informs clinical and public health decision-making.
9. Mathematical models of testing should make explicit their assumptions and sources of data; and investigate the impact of uncertainty. Estimation of the performance of test strategies of in vitro diagnostics requires empirical evaluation due to unknown sources of errors and likely oversimplification of modelling assumptions.

10. Planning for future pandemics should include:
    (a) Identification of multi-site networks to facilitate recruitment of patients or citizens willing to provide relevant biological specimens;
    (b) Creation, identification and maintenance of specimen banks;
    (c) Promoting active dialogue between public health, clinical medicine, laboratory medicine, statistical and methodological experts in test evaluation and regulators to agree on evaluation strategies;
    (d) Developing capacity and expertise in designing, delivering, analysing, and reporting studies of the clinical performance of tests in laboratory, clinical and community settings;
    (e) Expedited centralized processes for ethical and study-protocol approvals.

Box 2: Recommendations—Regulation matters.

(11) The Medicines and Healthcare products Regulatory Agency (MHRA) should review and revise the national licensing process for in vitro diagnostics to ensure public safety is protected, particularly in a pandemic. This review needs independent expert input from the relevant disciplines including appropriate statistical input.
(12) Scientific methods should be reviewed and developed to help regulators create Target Product Profiles that describe the characteristics and required performance of an in vitro diagnostic for a particular intended use.
(13) Regulators, in consensus with the scientific community, should specify reference standards judged to have acceptable accuracy against which the sensitivity and specificity of a new test can be established.
(14) Regulators' assessment of test safety needs to extend beyond the physical safety of a test device to the consequences of false positives and false negatives for those tested and all those affected by test outcomes. The full range of consequences, from liberalized behaviour to deprivation of liberty, should be considered.
(15) Evaluation of the impact of tests should ensure that both intended and unintended consequences are considered. Some consequences will not be evaluated before test implementation, so post-marketing surveillance for a new intended use requires ongoing assessment.
(16) During outbreaks, particularly when tests are being used outside their intended use, it is prudent to monitor test performance with regard to public safety, by requiring data collection and public reporting on: (a) test results, to assess whether a test is performing as expected in the target population; and (b) disease prevalence, to ensure tests are only used when they will do more good than harm.

Box 3: Recommendations -Transparency matters.

(17) Protocols for field or clinical evaluation studies should be publicly available to provide evidence of prior planning and to support transparency; and ideally should be prospectively registered.
(18) Expert peer review of study protocols and final reports by subject-matter (e.g. clinical, public health, laboratory) and methodology experts is recommended.
(19) Study reports should adhere to reporting guidance such as the Standards for Reporting Diagnostic Accuracy (STARD) to enable scrutiny of findings and incorporation in systematic reviews.
(20) Post hoc analyses should be limited; and clearly identified as exploratory.
(21) Study reports and results should be made available publicly in a timely manner.
(22) Field and clinical evaluation studies require ethical approval and fully informed consent as outlined in the Good Clinical Practice Guidelines.

(2002–2003), swine flu (2009–2011), Ebola (2014–2016), Zika (2015–2016), dengue (2016), plague (2017), and Covid-19 (2019–2023). Covid-19, the disease caused by SARS-CoV-2, emerged in China in December 2019 and is a contemporary example of the widespread and devastating impact of an infectious disease. The outbreak was declared a pandemic on 11th March 2020 by the World Health Organization (WHO).

Some infectious diseases are zoonotic, i.e. they can be transmitted from animals to humans. Zoonoses account for about two-thirds of human infectious diseases. Approximately three-quarters of emerging diseases in humans have originated in wildlife, including many of the most devastating pandemics in history such as the Justinian Plague (541–542 CE), the Black Death (Europe, 1347), yellow fever (South America, 16th century), and the Spanish flu (1918) (Machalaba et al., 2015). The Covid-19 pandemic appears to have zoonotic origins (World Health Organization, 2021).

The transmissibility of infectious diseases combined with age-related susceptibility or progression requires those infected to take precautions to prevent onward transmission to others and modify behaviours to reduce the risk of becoming infected. Personal protective equipment for healthcare workers may be needed to reduce their vulnerability to infection while they care for those affected.

Detection of infectious diseases is an essential part of combatting their spread, and the following sections outline some necessary key biological concepts.

2.1 Immune response

Pathogens have molecular components, known as antigens, which can trigger an immune response by the host's immune system to protect the body. A pathogen can have multiple different antigens which are unique to the pathogen. Lymphocytes, a type of white blood cell, are fundamental to the human immune system. The two primary types of lymphocytes are T cells and B cells. Immunity that occurs after exposure to an antigen from a pathogen or after immunization is known as adaptive or acquired immunity. The two types of adaptive responses are (1) cell-mediated immune response controlled by activated T cells and (2) humoral immune response controlled by activated B cells and antibodies in plasma cells.

Antibodies [immunoglobulins (Ig)] are protein molecules that can be found in blood and other body fluids such as mucus secretions and saliva. Antibodies are specific to a particular antigen: and bind to an antigen either to tag it for attack (binding antibodies) by white blood cells or to neutralize it (neutralizing antibodies). There are five types of antibodies—IgA, IgD, IgE, IgG, and IgM—with a range of functions. IgM is the first antibody the body produces in response to a new infection while IgG, the most common antibody, can take time to develop after infection or immunization. Once antigen-specific T and B cells have been activated following an infection, some cells persist, resulting in immunological memory for the specific antigens. During subsequent exposures to the same pathogen, the immune system is then able to mount a rapid and strong immune response to the antigens previously encountered. Some infections, such as chickenpox, induce a lifelong memory of infection. For other infections, such as seasonal influenza, immunological memory is less effective: the influenza virus evades neutralizing antibodies by regular mutation so that it is not recognized by antibodies that may have been produced in response to infection with a previous strain of the virus.

2.2 Stages of infectious diseases

The natural history of disease refers to the progression of a disease process in an individual over time in the absence of treatment (Centers for Disease Control and Prevention, 2023). Figure 1 illustrates a timeline from susceptibility to an infection to the end phase, culminating in recovery, disability, or death.

Transmission of infection can occur directly from person to person by physical contact, airborne routes via droplet or aerosol (e.g. SARS-CoV-2), fomite or indirectly (e.g. malaria parasites via mosquitoes). The stages of an infectious disease can be identified in terms of signs and symptoms of disease in the host (incubation and clinical disease) and the host's ability to transmit the pathogen (latent and infectious) (van Seventer & Hochberg, 2017). The interval between the time of exposure and the onset of disease symptoms (if symptoms appear) is known as the incubation phase.

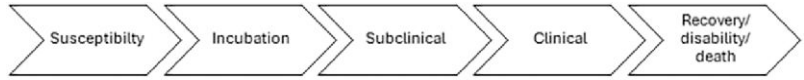


Figure 1. Natural history of disease.

This subclinical or asymptomatic phase can range from a few days (e.g. SARS-CoV-2 up to 14 days) to several years (e.g. HIV up to over 15 years). The next stage is the clinical disease phase during which signs and symptoms occur. Some individuals incubating an infection will not progress to clinical disease but may recover, have latent infection and be unable to transmit infection (e.g. latent TB) or act as carriers able to transmit infection to others (e.g. hepatitis B virus). For those who progress to clinical disease, the disease may be mild, severe or fatal (disease spectrum).

The infectious period, the period when an infected person can transmit the pathogen, depends on the disease, the pathogen, and the mechanisms by which the disease develops and progresses (van Seventer & Hochberg, 2017). For example, the infectious period for chicken pox is during the incubation phase while that of Ebola is during the clinical disease phase. Knowledge about the duration of disease stages is important for various reasons, including the appropriate use of testing in infection control and prevention strategies, defining the intended use of a test, and ensuring appropriate test evaluation.

Tests that can detect the pathogen or pathologic changes during the incubation phase when individuals are asymptomatic are useful for preventing the spread of infection, and also enable early intervention or preventive treatment. A key aim of testing during the clinical disease phase is for diagnosis to guide clinical management.

Influenza and hepatitis are examples that show that, even within the same family of viruses, infections can be transmitted differently; have different sequelae for those infected; different rates of ongoing infectiousness; different potential for control by immunization or treatment; and different consequences for the safety of donated blood or tissue. Protecting the blood supply from human immunodeficiency disease (HIV) galvanized innovation in the licensing of antibody and antigen tests for blood-borne infections.

Variant Creutzfeldt–Jakob disease (vCJD) exemplifies a dietary exposure potential epidemic that did not manifest in clinical cases to the extent feared. However, vCJD is also blood-borne, and the UK blood supply had to be protected despite there being no test for the abnormal prion protein that causes vCJD (see Supplementary Appendix).

For all these reasons, clinicians and researchers have devoted considerable attention to developing effective diagnostic tests for infectious diseases. The SARS-CoV-2 (the virus) and Covid-19 (the symptomatic disease) pandemic raised global challenges for clinical medicine, public health, economies, and everyday life. The development of new diagnostic tests is part of the necessary response. This report describes key issues in test evaluation of IVDs for infectious disease and investigates how they have been addressed in the evaluation of new tests for SARS-CoV-2.

3. Diagnostic tests

Diagnostic tests can indicate the presence or absence of infection or a surrogate marker, or detect evidence of previous infection (e.g. antibody tests). An infected individual may show signs and symptoms or may be asymptomatic. For some infections like SARS-CoV-2, asymptomatic individuals may transmit infection to others. Therefore, for both symptomatic and asymptomatic infections, early identification is often essential for effective clinical and outbreak management, including the implementation of control measures such as contact tracing to interrupt transmission.

Diagnostic tests for infectious diseases have multiple uses: patient management, screening for asymptomatic infection, surveillance; evaluating the effectiveness of interventions (including vaccines and verification of elimination), and detecting infections with markers of drug resistance (Banoo et al., 2006). Ongoing scientific advances lead to the development of new diagnostic tests for improved management and control of infectious diseases. The reported performance of a

diagnostic test is not an inherent property and can be influenced by factors such as the characteristics of the population and infectious organism, test format, technical expertise, and study methods. Therefore, tests should be rigorously assessed in appropriate laboratory, clinical, and/or field settings to ensure their validity and applicability in practice.

The focus of the RSS Working Group was on IVDs, which are the most common type of test for diagnosing infectious diseases. However, many of the issues raised are generic and apply to other test types, such as imaging.

3.1 Types of in vitro diagnostic tests for infectious diseases

In vitro diagnostics are tests done on samples such as fluids or tissue that have been taken from the human body. IVDs can detect diseases or other conditions and can be used to monitor a person's overall health to help cure, treat, or prevent diseases (Food and Drug Administration, 2023). See Box 4 for descriptions of different types of IVDs.

Box 4: Types of in vitro diagnostic tests for infectious disease.

Microscopy

Microscopy enables the visualization of a pathogen by using a microscope to examine a specimen (e.g. blood or tissue). To enhance contrast, specimens may be treated with stains to colour certain features of the pathogens or the background. The choice of stain will depend on the pathogen, e.g. Giemsa stained thick or thin blood smear for detecting malaria parasites. Wet mounts (i.e. a drop of liquid on a slide) of unstained specimens can be used to detect pathogens such as fungi. Microscopy is useful for species identification and quantification.

Culture

Pathogens can be cultured under controlled laboratory conditions in artificial nutrient mediums such as nutrient broths and agar plates. Unlike most bacteria that can grow in artificial media, viruses require a living host cell for replication. Virus culture can be achieved either through a living host, an embryonated egg, or tissue/cell culture. Microbial cultures can be used to identify the type of pathogen, the quantity in the sample, or both.

Molecular tests

Molecular tests detect a pathogen by measuring specific genetic sequences in deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) or the proteins they express. An essential process underpinning many molecular diagnostics is amplification. This process makes copies of a specific DNA or RNA sequence found in a sample until there are so many copies that they can be detected and measured. There are several amplification techniques but the most commonly used is gene amplification by polymerase chain reaction (PCR). Rapid molecular tests can improve the time to diagnosis and access to testing, e.g. the Xpert MTB/RIF assay recommended by the WHO as an initial test for diagnosis of tuberculosis (TB) or rifampicin-resistant TB.

Serology

Serology or antibody tests are blood-based tests that check for an immune response to identify individuals who have had a particular infection or developed immunity. Blood samples are tested for antibodies to the pathogen by combining the samples with specific antigens of the pathogen. If the antibodies are present, they will stick to the antigens. The body does not

produce antibodies against a new pathogen immediately and so antibody tests cannot detect such infections at an early stage.

Antigen detection
Antigen tests detect the presence of an antigen, and so can detect current infection with a pathogen. Antigen tests can take longer to develop than molecular and antibody tests due to the need to first identify suitable antibodies for the assays. Antigen tests are amenable to point-of-care use, thus making them more suitable for testing in the community and in remote settings, e.g. rapid antigen tests for diagnosis of malaria.

Drug resistance tests
Drug resistance leads to increased morbidity and mortality; and presents a great challenge to disease control. With the increasing use of antimicrobial drugs, antimicrobial drug resistance has become a major clinical problem. Culture-based phenotypic susceptibility testing and molecular diagnostics are frequently used to detect drug resistance.

3.2 Setting of testing

Laboratory testing can be performed in different settings, from centralized laboratories to self-testing in homes. Laboratory tests that are performed at the point of care, e.g. at the bedside or in a clinic rather than in a laboratory are often referred to as point-of-care tests (POCTs). Ehrmeyer (Ehrmeyer & Laessig, 2007) defined POCT testing as 'patient specimens being assayed at or near the patient, with the assumption that test results will be available instantly or in a very short time frame, to assist caregivers with immediate diagnosis and/or clinical intervention'. Technological advances have led to innovation in portable devices that are easy to use and can give results within a shorter timeframe compared to tests performed in conventional laboratories. Accurate rapid diagnostic tests (RDTs) for infectious diseases can revolutionize patient management and disease control through increased diagnostic capacity, quicker turnaround times and improved accessibility, particularly in resource-limited settings. Such RDTs have been endorsed by the WHO for diagnosis of TB and malaria, two life-threatening infectious diseases with significant global disease burden. These RDTs may be POCTs or laboratory-based tests.

3.3 Challenges in testing for infectious diseases

Most infectious diseases are caused by viruses, bacteria, or parasites. In the Appendix (Supplementary Material), pandemic and seasonal influenza (viral), hepatitis (viral), HIV (virus), vCJD (abnormal prion), TB (bacterial), malaria (parasitic), and Covid-19 (viral) are used to highlight some of the key challenges of infectious diseases which impact on the development and evaluation of diagnostic tests. These include having a clear definition of the target condition to be detected and the population in whom the test will be used, the nature and intended use of the test, and the availability of an acceptable reference standard for verifying the presence or absence of the target condition.

4. Statistical parameters required for reporting

Test evaluation from bench to bedside is multifarious. Throughout this report we refer to the test under evaluation as the index test—it may be a new test or an existing test being considered for a new purpose. Horvath et al. (2014) identified five key components of test evaluation which assess the different steps required to evaluate an index test: (1) analytical performance, (2) clinical performance, (3) clinical effectiveness, (4) cost-effectiveness, and (5) broader impact. Analytical performance refers to the ability of an index test to conform to predefined technical specifications whereas clinical performance assesses whether a test detects patients with a particular clinical condition or in a physiological state. Clinical effectiveness refers to the ability of a test to change health outcomes that are relevant to the individuals being tested (also referred to as assessments of clinical utility or impact, and can include both benefit and harm). Cost-effectiveness compares the impact

of tests with the resources that they use. Cost-effectiveness and clinical effectiveness are frequently addressed jointly. This joint approach to effectiveness is known as societal efficacy when costs are considered from a societal perspective (Takwoingi, 2016). Finally, broader impact refers to all other consequences of testing beyond clinical and cost-effectiveness (Horvath et al., 2014). Several frameworks have been proposed to map the different steps in this process (Lijmer et al., 2009), which tends to be cyclic and repetitive rather than linear.

Section 4 focuses on the first two components: analytical and clinical performance. Studies of analytical performance (Section 4.1) provide only the most basic demonstration that the test can work in optimal laboratory conditions. Less well understood and sometimes ignored is the need to demonstrate clinical performance (Section 4.2) before the test can be recommended in a target population. A brief summary of issues in clinical effectiveness studies (Section 4.3) introduces the concepts of evaluating the impact of an index test on consequent health outcomes and other mechanisms by which they affect patients.

4.1 Analytical performance

Studies of the analytical performance of a new test are performed in controlled laboratory settings to establish the measurement properties of the assay under ideal conditions. Studies of analytical capabilities and performance provide evidence of the measurement properties of a test, indicating how well it can detect and/or correctly measure the pathogen or relevant biomarker (e.g. molecule, antibody, antigen, or other; see Box 4 Section 3.1). Analytical performance assesses whether the assay can deliver basic quality specifications that are required for the test to have the potential to be a usable detection mechanism for the infection (present or past).

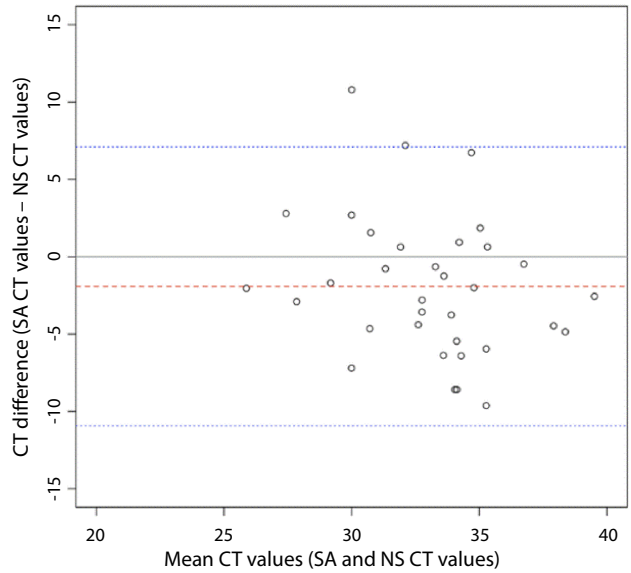
Studies to evaluate analytical performance are first carried out by manufacturers and subsequently must be repeated in independent laboratories with the objective of verifying analytical performance claims reported by test developers. Several measurement properties are typically considered, such as imprecision, bias, reproducibility, clarity of test operation, and clarity of results interpretation. Of these, imprecision and bias are central to the initial determination of a test's potential value and essential to establish for regulatory purposes. Detailed documentation on these quantities and processes by which they are estimated are summarized by the Clinical and Laboratory Standards Institute (CLSI) (https://clsi.org) and the International Organization for Standardization (ISO) (www.iso.org). This sub-section briefly summarizes key aspects of the assessment process to give context but refers readers to a wealth of detailed documentation at CLSI and ISO for full explanations. The CLSI Harmonized Terminology Database (https://htd.clsi.org) may be of particular help as terms are used in peculiarly precise ways in analytical studies which are often different to the way the same terms are used in clinical studies (for example, the word accuracy).

4.1.1 Imprecision

Imprecision, sometimes referred to as repeatability or precision, quantifies the impact of random variation on how likely repeated observations from the same sample are to provide (theoretically) similar results. Imprecision for numerical biomarkers is most often summarized as a coefficient of variation (CV), defined as $CV = (\text{Standard deviation}/\text{Mean}) \times 100$. Other metrics, such as reference change values (RCV), use the CVs for laboratory variation and within-individual variation as a guide to the significance of observed changes over time. Biomarkers with high CV or RCV values are only appropriate for determining large differences and are unlikely to deliver high diagnostic accuracy.

4.1.2 Bias

Bias, which directly relates to test accuracy or 'degree of trueness' (Johnson, 2008), measures how closely the average of a set of measurements agrees with the 'true value'. Studies are often designed to compare results per laboratory when testing external specimens from a quality assurance scheme or from a national standard laboratory. The CLSI suggests 20 specimens that span the range of interest (Carey et al., 2005). Correlation, which measures scatter (imprecision) and has nothing to do with agreement, has frequently been misused to assess bias. A difference plot (difference against known value), as suggested by Bland & Altman (1995), is preferred (see



Footnote: Differences were calculated between CT values of SA and NS samples. The red line represents the average difference between sample types, and the blue lines represent estimated limits of agreement.

Figure 2. Bland–Altman plot of SARS-CoV-2 CT values for saliva (SA) and nasal swab (NS) samples (excerpt from Grijalva et al., 2020; distributed under creative commons attribution license).

Figure 2). When the differences are between known values and measured values, the distribution of differences provides an estimate of bias; and plotting against a known value allows assessment of whether the bias is constant or associated with the measurement levels. When a pair of devices is being compared, the distribution of differences is better plotted against the mean.

4.1.3 Analytical sensitivity

The term analytical sensitivity refers, on the one hand, to the ability of a test correctly to classify biological samples as positive and, on the other, to the detection capability of a test, based on three limits: limit of blank (LoB), limit of detection (LoD), and limit of quantification (LoQ). These limits, described in the CLSI guideline EP17 (Tholen et al., 2004), *are based on a specific analyte* (*a* substance used for identification and measurement).

The definition of these quantities in the standard text is based on an assumption of normally distributed measurements. The LoB measures the 'highest apparent analyte concentration expected to be found when replicates of a blank sample containing no analyte are tested' (Armbruster & Pry, 2008). By contrast, LoD measures the 'lowest analyte concentration likely to be reliably distinguished from the LoB and at which detection is feasible'. The assumption that observations are normally distributed on the chosen measurement scale may be particularly challenged by large numbers of observed zeros when estimating the LoB. The LoQ is the 'lowest concentration at which the analyte is reliably detected and at which predefined goals for bias and imprecision are met'. Most of the discussion in product leaflets for IVDs for infectious diseases concerns LoD, but all three are related and relevant. In particular, LoQ could be the same as LoD but, in some instances, may be significantly higher.

4.1.4 Analytical specificity

Some tests may pick up a range of pathogens, which are false positive (FP) findings if they are not the condition of interest. Analytical specificity assesses whether the assay is likely to give these FP results and is assessed by using the assay on stored samples from persons known to have other conditions, or spiked samples. The list of conditions against which the assay is assessed needs to

include those which are most likely to give FP findings, and most likely to present in a similar way. There does not appear to be a standard process for defining this list.

It is also important to assess whether certain conditions interfere with assays (e.g. rheumatoid factor, bilirubin, lipids) and to assess whether the ability to detect disease is compromised, yielding false negative (FN) results.

4.1.5 Sample size requirements for analytical performance studies

Suggested sample sizes are included in documentation from CLSI and ISO, but their statistical basis is often unexplained. The LoB is estimated by measuring replicates of a blank sample and calculating the mean value and standard deviation (SD). A sample size of 60 is suggested for establishing the LoB, with sample sizes of only 20 for local laboratory verification. The LoB is estimated as the 95th percentile value of a normal distribution with the calculated mean and SD.

Many assays are not able to measure the smallest concentrations or LoD. Two approaches are discussed in the literature to obtain this measure. The first is based on the distribution of apparent analyte concentrations in replicates of blank samples, as for LoB, but using a higher percentile value from the assumed normal distribution, at 2, 3, 4, or even 10 SDs above the mean, although this approach is now regarded as invalid. Alternatively, a more widely recommended empirical approach makes use of analyses of samples with known low concentrations of the analyte of interest.

4.2 Clinical performance

4.2.1 Clinical studies

There are two distinct designs of studies which produce estimates of clinical sensitivity and clinical specificity (sometimes known as diagnostic sensitivity and diagnostic specificity). The first type of study applies the index test in samples from two pre-selected groups of people already *known to have the target condition* and *not to have the target condition* which can be accessed and tested quickly and efficiently. Such studies (referred to as two-group or two-gate designs, or diagnostic case–control studies (Rutjes et al., 2005)) often use existing samples tested in laboratory settings. Some studies may either look at performance in those known to have the target condition (and only estimate sensitivity), or in those known not to have the target condition (and only estimate specificity). The selection of the groups will influence the estimates and needs to be fully described and justified. Unsuitable selection may lead to bias.

The second type of study is field studies that evaluate the performance of the test in a real-world setting, applied to people/patients for the test's intended use in clinical practice and assessed against a reference standard. Participants are recruited as a single group before it is known whether they do or do not have the target condition. Thus, all patients considered suitable for the test are included. The distinguishing feature is that participants are representative of individuals for whom the test would be used in practice and are recruited prior to their disease status being ascertained.

Differences have been observed between the results of studies which adopted these two designs: those using preselected patient groups having estimates of clinical performance that are higher than those observed when tests are undertaken for their intended use in real-world settings (Lijmer et al., 1999). This bias may occur as individuals who are the most difficult to diagnose (and most likely to give FP or FN results) are often excluded from two-group studies, as only those with known disease states can be recruited.

As both designs estimate sensitivity and specificity, it is essential to be aware of how participants were selected in each study when interpreting its results. These issues are discussed further in Section 5.2.2.

4.2.2 Diagnostic or clinical sensitivity and specificity

Diagnostic or clinical sensitivity and specificity describe the test's performance in terms of the proportion of individuals with the target condition who are correctly detected by the index test and the proportion without the target condition whom the index test correctly identifies as negative. Values of clinical sensitivity and specificity are not fixed constants but vary with context and intended use.

Studies which evaluate the accuracy of two or more tests will also report estimates that compare sensitivity and specificity, or FP rate (computed as 1-specificity) and FN rate (computed as 1-sensitivity), either as ratios or absolute differences.

4.2.3 Predictive values

Sensitivity and specificity depend on context of use, but predictive values depend additionally on the prevalence of the target condition. Predictive values are statistics that explain the probabilistic meaning of positive and negative test results. The positive predictive value (PPV) is the proportion of individuals receiving positive test results who have the target condition. The negative predictive value (NPV) is the proportion receiving negative test results who do not have the condition. As positive and negative predictive values mathematically depend on prevalence, their estimates should be based on clinical sensitivity and specificity and on a range of infection prevalence, as the latter is rarely adequately estimated or modelled. Studies should be carried out in settings as close to the test's intended use as possible; and not from bias-prone two-group studies (see Section 4.2.1). Presenting the impact of varying prevalence helps to determine the potential utility of estimates of predictive values (PPV and NPV).

4.2.4 Indeterminate results and test failures

Besides clinical sensitivity and specificity, a field evaluation should report how often test results are inconclusive or could not be obtained: the void or invalid rate gives an indication of the suitability of the test (Shinkins et al., 2013). Therefore, reporting the number of inconclusive and missing results for all tests (including the reference standard) is critical: such as in a table giving all results as part of a clinical agreement study (Food & Drug Administration, 2007).

4.2.5 Clinical effectiveness and other aspects of test performance

The clinical impact of tests on health depends on the consequences of the interventions that follow, which are best assessed through randomized controlled trials (RCTs) comparing different intervention strategies and the consequential patient outcomes (Bossuyt et al., 2000; Fryback & Thornbury, 1991). When tests give false results (FNs and FPs), the actions which follow may cause harm. For example, in a pandemic, the use of tests which falsely state that individuals are not infectious can increase disease spread.

Aspects of the tests, other than their performance, impact their utility in a given setting. For example, timing (e.g. how long it takes from obtaining a biological sample to test result) and human factors that impact usability and user errors need to be investigated (e.g. handling of machinery, collection devices, specimens, etc.) either as part of the evaluation of clinical performance or before evaluating clinical effectiveness. Results for some of these issues will be captured as invalid and missing results, but full disclosure of the reasons for missing results can be important.

5. Study designs for clinical studies

Key aspects of clinical performance studies involve specifying the intended use or purpose of the new (index) test (Section 5.1) and the ideal study design (Section 5.2). Section 5.3 considers the implications of variations from the ideal study, and Section 5.4 addresses sample size. Brief summaries of other clinical study designs are described for clinical effectiveness studies (Section 5.5), studies of natural history (Section 5.6), and surveillance (Section 5.7).

5.1 Specifying the intended use or purpose

Specification of the intended use or use case is key to the appropriate evaluation to inform testing policy (Doust et al., 2021; Horvath et al., 2014). An intended use case describes the application of a test in a particular patient group to diagnose a stated condition: namely, the who, where, when, what, how and why a test is applied. Part of the test's intended use is the role the test is likely to play within the clinical pathway. The most common roles for a test are for triage use of a confirmatory test, as an add-on or a replacement for existing tests, or as a new test which opens up a completely new treatment pathway (Korevaar et al., 2019). The same test can perform different roles in different clinical settings and pathways.

Clinical performance studies involve assessing the sensitivity and specificity of a test for its intended use and in participants similar to the target population who will be tested in practice. This means recruiting representative participants in the right setting, testing at the intended time point, obtaining the required specimens and testing them as would be done in practice. Alongside this, a reference standard for diagnosis, which is the closest possible to the truth, must be obtained. The reference standard may be one or more tests (undertaken independently of the index test) with or without additional clinical information obtained subsequently. The results of the index test and reference standard are then cross-classified to compute sensitivity and specificity.

Too often, the intended use of a test is not recognized as a critical aspect of its evaluation. For example, it is essential to recruit a group representative of those in whom the test will be used, as test performance will change with the spectrum of the severity and stage of the disease (for sensitivity) and the competing conditions from which it must be distinguished (for specificity) (Ransohoff & Feinstein, 1978). In addition, aspects of how a test is undertaken in real life may reduce its performance. In *extremis*, even for a test with perfect analytical sensitivity and specificity, the practicalities of delivering a test in a real clinical setting may make its use infeasible. Once the intended use for the test is clear, laboratory evaluations using appropriate samples from a representative population may be necessary to determine feasibility before embarking on large-scale field studies, highlighting the cyclic nature of the evaluation process.

There can be multiple intended uses for a test, see Section 3. The setting (e.g. laboratory, hospital, general practice, surveillance-location, home use) can be used as a simple proxy for understanding how far removed from the laboratory environment the setting is in which the tests are being used. Stating where the test is meant to be used and evaluating the test in that setting would be the minimum expected to determine evidence of clinical performance.

5.2 Study design considerations

The basic study design for a clinical study involves prospective recruitment of a representative sample of patients/participants from the target population, identified as those on whom the test will be used in clinical practice for its intended use (Doust et al., 2021). The index test and reference standard are carried out on all participants, with all tests done at the same time or within a time interval deemed acceptable given the nature of the target condition. The execution of the study must be undertaken to minimize the risk of bias, for example by ensuring that the index test and the reference standard are undertaken and read independently of each other, and that complete data are obtained. Key elements depend entirely on the intended use of the test; we expand on each of these below.

5.2.1 Target population

The performance of an index test depends on the population on which it is tested. The intended use will define the key characteristics of the study population that determine test use. Whilst participant characteristics such as age, sex and ability to provide informed consent are commonly required, the most critical criteria are those that determine the clinical reasons that trigger testing, or the population characteristics that make someone eligible for screening.

Inclusion and exclusion criteria should list the signs and symptoms that ensure those recruited will be representative of the population that eventually will be targeted for the use of this test. These symptoms by themselves already increase the likelihood of infection compared to those that are asymptomatic, which impacts on the proportion of overall positives (for both the index and reference tests) and therefore on the proportion of positives that are FP compared to true positive (TPs). However, they may also impact on the stage and severity of the disease which will directly affect sensitivity.

5.2.2 Prospective recruitment of a representative sample of participants

A representative sample can be obtained by prospectively recruiting consecutive participants in the clinical setting and pathway in which the test will be used. (Whilst random samples theoretically will also yield representative samples, they are rarely possible in clinical settings). Retrospectively collected data are likely to introduce selection bias, for example by focusing on those that received

an intervention (e.g. hospitalization) because they exceeded a decision threshold (e.g. based on severity of disease) and so do not represent the target population for whom the test was designed. At the same time, location will play a significant role in the identification of the study population; patients attending general practice may differ from those seen in hospital. The prevalence of infection will be different, as well as the characteristics of the individuals, with potential differences in the symptoms observed in each setting (disease spectrum) (Holtman et al., 2019).

5.2.3 Timing of the tests

Whatever the test is actually measuring (e.g. pathogens, molecules, antibodies, antigens, etc.) may be unlikely to remain stable over long periods of time (there are exceptions such as HIV antibodies). This could be due to the natural history of the disease, which includes the body's immune response or, in some cases, interventions such as drug treatments. Given this, the timing of the test in relation to, for example, the first symptom can have a substantial impact on the ability of the test correctly to discriminate between those who have and those who do not have the infection. Too early in the infection, the level of antibodies or antigens might be too low for the test to be able to identify this. Similarly, if it is too late the levels might again have reduced to undetectable levels. This is the main reason that a diagnostic accuracy study for acute infection is generally expected to be designed so that the index test and the reference standard are performed on specimens collected at the same time, thereby providing a fair comparison.

5.2.4 Delivery of tests

Clear protocols should not only describe the target assay but also specify test delivery: who will obtain the specimens (e.g. nurse, self-sample); what will be sampled (e.g. blood, saliva); how will the specimen be obtained (e.g. finger prick, spitting into a tube), stored and transported; and when (e.g. within 7 days of first symptom). These elements should clearly relate to the proposed instructions for use (IFU) of the test so that they reflect the way the tests will be used in practice. Within this protocol, information about how blinding the results of the tests and the reference standard is carried out is also necessary to guarantee unbiased results (see Section 5.3.2).

5.2.5 Comparisons of tests

It is common that multiple tests are developed to identify the same target condition. Studies that carry out head-to-head comparisons, for example, using specimens from the same individuals, are particularly useful to identify potential differences in clinical sensitivity and specificity, as well as other aspects of test performance (Takwoingi et al., 2013). The basic study design is still based on the principles described in this section but with two or more diagnostic tests included instead of only one (in addition to the reference standard). This is sometimes referred to as a paired or within-person design. Alternatively, when it is not possible to collect specimens from the same individual to evaluate multiple tests, a randomized design can be used whereby participants are randomly allocated to one of the tests, but all participants receive the reference standard. In either design, particular care is required to ensure that the timing of tests is arranged to minimize potential bias. For example, if multiple specimens are required for the different tests and can only be taken separately, randomizing the order these specimens/tests are taken would likely prevent systematic bias (see Section 5.2.4).

5.3 Variations in study design—impact on validity (internal and external)

Some reasons that the ideal study design cannot be achieved include: critical urgency in determining basic levels of accuracy, extremely low prevalence of the infection which makes recruitment of consecutive participants potentially wasteful (not enough cases), carrying out the study in the relevant setting is extremely difficult, and specimen required for the reference standard is not feasible in all participants (because invasive).

Empirical evidence has shown which variations are most likely to affect a study's validity (Whiting et al., 2011). This evidence was taken into account in creating the Quality of Assessment of Diagnostic Accuracy included in Systematic reviews tool (QUADAS-2) (University of Bristol, 2011) used in systematic reviews to assess the risk that a study's findings may be biased and may not be directly applicable to the intended use case. The tool organizes

its considerations in the following four domains: (1) patient selection, (2) index test(s), (3) reference standard, and (4) flow and timing (covering completeness of data and standardization of verification and timing).

5.3.1 Could the selection of patients have introduced bias?

The key consideration is whether the study's sample of patients is representative of those for whom the test will be used in practice.

Complete recruitment of a consecutive series of participants is often not feasible, as participants can only be recruited if they consent and when study staff are available. Whether such restrictions introduce bias will depend on the degree to which those recruited differ systematically from those who are not.

Two-group/diagnostic case–control designs, wherein individuals are recruited from different groups already known to have and not have the infection, routinely fail to recruit representative samples (see Section 4.2.1). Bias occurs because these two groups, which have already been adequately differentiated, typically over-represent those with severe disease and those completely free of all disease, while those with uncertain status are usually excluded. Hence, individuals in each group are likely to be in the extremes of their distribution and could, therefore, artificially aid the performance of the test. This bias will be reflected in an over-optimistic estimation of the performance of the index test.

Similarly, over-sampling and under-sampling particular groups leads to bias in overall estimates of sensitivity and specificity unless done with properly structured probabilistic sampling which has to be accounted for in the analysis.

5.3.2 Could the conduct or interpretation of the index test have introduced bias?

Interpretation/measurement of index and reference tests must be independent, i.e. results for each should be obtained blinded to the other. If the reference standard result is known in advance, this can potentially affect the interpretation of the index test, particularly where the index test provides an inconclusive or borderline result. It could also lead to a re-interpretation of results and the retrospective evaluation of why there is a disagreement between the index test and the reference standard.

In a diagnostic accuracy study, it is expected that the characteristics of the index test, including the threshold used to define test positivity (positive case), should be predetermined during the development of the test. Using the information from the study to define the threshold will lead to an overestimation of the index test's performance.

5.3.3 Could the conduct or interpretation of the reference standard have introduced bias?

It is important to use reference standards which provide the most accurate classification of participants possible, but often the reference standard is not a perfect classifier. This means that the reference standard itself makes errors leading to the incorrect classification of some correct index test results as FPs or FNs. The use of multiple measures and composite reference standards may improve classification (Glasziou et al., 2008; Naaktgeboren et al., 2013). Where no suitable reference standard exists at all, accuracy studies may not be possible, and it will be important to assess the impact of the test on diagnostic and treatment decisions and, ultimately, patient outcomes (see Section 5.5).

As described in 5.3.2, adequate evaluation of the index test against the reference standard relies on their independence.

If, in the extreme, knowledge of the index test's result determines whether to carry out the reference standard or if the reference standard is changed when it disagrees with the index test (discrepant analysis (Hadgu, 1999)), substantial bias is likely, typically artificially increasing the estimated accuracy of the index test.

5.3.4 Flow and timing—could the patient flow have introduced bias?

It is important that the new and reference tests are undertaken close enough in time for there to be little chance of there being any change in patients' disease status between the tests. The ideal scenario assumes that the new and the reference tests are performed at the same time, but this may not

be possible for logistical reasons: for example, if different specimen types are required, which necessitate repeat visits. Even when both tests require the same specimen type (e.g. both require nasopharyngeal swabs), the specimen could be reduced substantially by the time the second swab is taken, which biases against the second test. In such situations, randomizing the order of within-person tests will be necessary.

There are some situations where there are multiple accepted reference standards and the design allows for more than one to be used to evaluate an index test. This could be because it is not possible for all patients to receive the same reference standard (e.g. if this requires a certain duration of follow-up). In these scenarios, it is important to highlight which reference standards were used and on which participants and mention potential advantages/disadvantages of using each reference standard.

As discussed in Section 5.2.1, patient selection is critical as the exclusion of certain participants is likely to generate bias. For the same reason, analysis of all available participants is necessary with clear explanations, justifications, and discussion as a potential limitation, whenever this is not feasible.

5.4 Sample size issues and incorporating or evaluating uncertainty

The choice of methods and approaches to sample size estimation for studies assessing clinical performance will vary depending on the objectives (Obuchowski, 1998). For example, if the regulatory focus specifies a required performance (e.g. expected sensitivity of 90% with a minimum performance requirement of 80%) then, based on expected performance and estimates of prevalence, study size can be determined so that, with high probability, the lower limit of the 95% confidence interval for sensitivity exceeds the minimum performance required.

A similar approach can be used focusing on required precision around an estimate while varying the expected performance of the test as well as the prevalence. This approach helps inform the maximum number of participants required, which is usually viewed as conservative.

When the focus is on hypothesis testing and/or direct comparisons, sample size estimation similar to that for randomized trials is normally carried out based on expected differences in parameters (e.g. sensitivity or specificity), probability of type one error and power (alpha and 1-beta respectively). Korevaar et al. (2019) describe a clear framework for such studies and they provide a file for the calculation of sample sizes based on this approach. Sample size methods for paired samples are required in studies where each individual receives multiple index tests (Alonzo et al., 2002).

Of note, reviews that have explored the reporting of sample size estimation in diagnostic studies have identified that most do not report any formal calculation (Bachmann et al., 2006; Bochmann et al., 2007; Thombs & Rice, 2016). Reporting of sample size in diagnostic studies was only included in the 2015 version of STARD (Bossuyt et al., 2015) and so it is possible that the proportion of studies adequately reporting formal sample size calculations has improved since then.

5.5 Clinical effectiveness studies of the impact of tests and testing strategies

Given that the impact of testing for infectious diseases to reduce transmission depends entirely on the consequent behaviour of individuals, evaluation of the broader impact of testing encapsulates both accuracy and nonaccuracy impacts, which can affect both the benefits (often counted as cases detected) and harms (for example, unnecessary isolation, disinhibition from FN test-results leading to the potential for increased transmission, lost income). New tests, particularly POCTs, may radically change access to testing, leading to differences in those who gets tested, which needs to be accounted for in evaluating their impact.

Frameworks such as the *Ferrante BMJ Framework* have detailed the routes by which tests impact on patient outcomes (Ferrante di Ruffano, Davenport et al., 2012; Ferrante di Ruffano, Hyde et al., 2012) via intended and unintended effects—and can help in planning evaluations which need to be undertaken. Effects can be categorized under four main headings: (a) direct test effects on the patient (e.g. risk of harm, procedural discomfort and anxiety, and reassurance); (b) altering clinical decisions and actions (related to correct use of test and interpretation of the test result); (c) changing time-frames of decisions and actions (e.g. reducing time to diagnosis and treatment); (d)

influencing patient and clinician perceptions and behaviours (e.g. willingness to undergo procedures, the impact of test results on patient behaviour, defensive medicine).

Large cluster RCTs comparing different testing options can be difficult logistically or in policy terms, but they may be necessary to ensure clinical effectiveness and value for money. Randomized step-wedge designs may be the only option if a decision has already been made to roll out a test policy. Accumulating portfolios of evidence using studies of different designs may be a faster way to capture the breadth of positive and negative impact that testing can have.

5.6 Studies of the natural history of disease

One important example is studies that aim to determine the appearance (seroconversion) and persistence (decay) of antibodies post-infection. Information from these studies is particularly relevant to understanding the natural history of the disease post-infection and determining the timing of potential tests that are based on antibody-level detection. The ideal design for these studies will be based on individuals whose initial infection date can be ascertained and from whom repeated measurements (typically blood samples) are taken over an extended follow-up period (e.g. a longitudinal study) (Iyer et al., 2020). Variations from this design, such as uncertainty in the timing of infection or the use of multiple cross-sectional samples of participants instead of acquiring longitudinal data, can be considered to minimize the potential for bias.

5.7 Surveillance studies

Surveillance studies of infectious diseases have attempted to quantify:

(a) exposures (consumption of bovine spongiform encephalopathy (BSE) contaminated foods; sexual attitudes and lifestyles; injection drug use; contacts—who meets whom; mobility);
(b) prevalence (or incidence) in risk groups (antenatal women; new-born babies; patients at genitourinary medicine clinics; injection drug users; prisoners; healthcare workers);
(c) prevalence (or incidence) in potentially nationally representative groups (blood donors; persons undergoing appendectomy; individuals on NHS register; community-living members of households).

Surveillance studies usually require linkage of different data sources. Three types of surveillance study which link a biological specimen (to be tested) with brief demographical and/or exposure information about the person who gave the specimen are described below.

5.7.1 Unlinked anonymous testing

Unlinked anonymous testing (UAT) uses a residue or aliquot from a testable biological specimen given for other reasons. It requires ethical approval but is unconsented by individuals. Individuals may opt out, as information about UAT is posted in clinics or blood donation centres, as appropriate. These UAT studies are designed so that there can be no deductive disclosure about individuals; hence, only minimal information about the person whose specimen is tested (such as sex, broad age group, and region) is retained with the surveillance specimen.

5.7.2 Consent with attributable linkage of the biological specimen (to be tested) and brief risk factor questionnaire or interview

High volunteer rate matters, as does representative sampling. Volunteers expect to be notified about their individual test result and they take part having been assured about the confidentiality of their linked test result and risk factors. Consented attributable linkage, or the third surveillance option, is necessary when the biological specimen (to be tested) is not routinely stored (e.g. nasopharyngeal swab in SARS-CoV-2).

5.7.3 Consent for nonattributable linkage of biological specimen (to be tested) and brief self-completion risk-factor questionnaire

Volunteers understand that the linking of their biological specimen and risk-factor questionnaire is done in such a manner that the linked-pair is not attributable to the individual to whom they

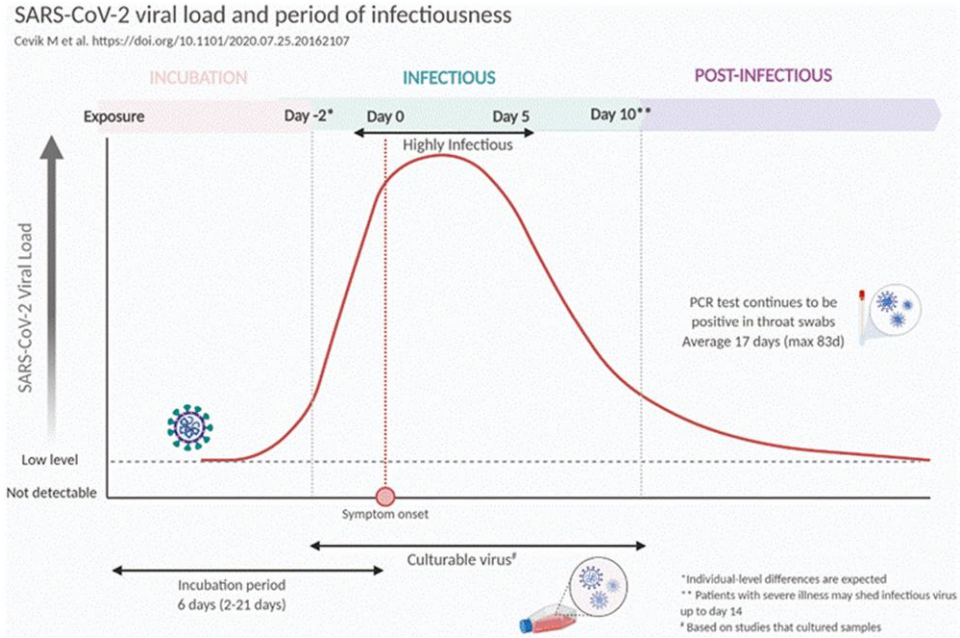


Figure 3. Schematic of viral levels during a COVID-19 infection (excerpt from Cevik et al., 2021; subject to copyright permission from Oxford University Press).

belong; and hence that individual test results cannot be reported back. However, the results for their community (prison, school, accident & emergency department or antenatal clinic) shall be reported-back (Bird et al., 1992; Gore et al., 1995, 1999; Hutchinson et al., 2000; White et al., 2015; Yirrell et al., 1997). High volunteer rate matters, which nonattribution encourages so long as results are reported back to the community and the research team has ensured the community's easy access to confidential testing on a personal basis. Self-completion questionnaires about risk behaviours afford privacy and engender frankness.

6. Lessons learned from evaluating tests during COVID-19 pandemic

6.1 Introduction

Tests used during the Covid-19 pandemic are of two main types: tests which detect the virus or parts of the virus (molecular and antigen tests) and tests which detect immunological response to infection with the virus (antibody tests) (see Section 3.1). Both test types are used for multiple purposes, in different groups of patients and citizens, and at different time points.

The deployment and performance of tests require a basic understanding of the kinetics of both the viral infection and the antibody response and the heterogeneity which may be observed in these patterns between individuals. As with all infectious diseases, after infection, viral levels rise to a peak as the virus proliferates and then subsequently fall as the immune system responds (see Figure 3 (Cevik et al., 2021)). Initial immune responses (see Section 2.2) are of IgA and IgM antibodies, with IgG appearing later and lasting longer (see Figure 4 (Post et al., 2020)). The key, however, is understanding how the timing and magnitude of these rises, peaks, and falls relate to patient characteristics, exposure, onward transmission, and symptoms, as this determines the roles that tests can have for early detection, diagnosis and surveillance. At an early stage of the pandemic, knowledge of these details may be limited, but it is important that the consequences of such limitations are made clear when introducing and evaluating the tests.

Our knowledge of these trajectories has been acquired through longitudinal observation of both small and large patient groups (Gudbjartsson et al., 2020; Hall et al., 2021), opportunistic epidemiological studies of communities (such as outbreaks on cruise ships (Hung et al., 2020;

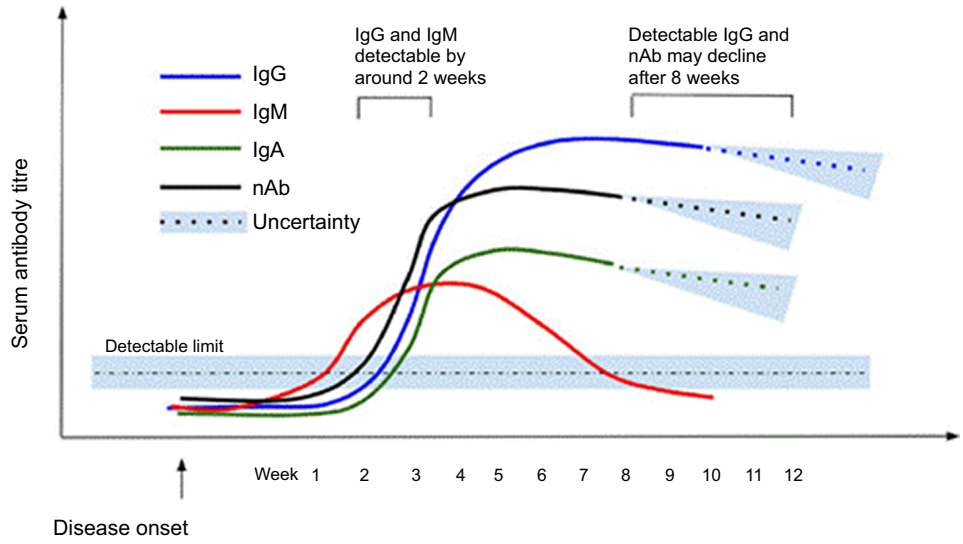


Figure 4. Schematic of antibody levels during a COVID-19 infection (excerpt from Post et al., 2020; distributed under creative commons attribution license). The y-axis is illustrative so no scale shown.

Mizumoto et al., 2020)) and through insights gathered via cross-sectional studies using diagnostic tests (Deeks et al., 2020). One initial challenge was to foresee and undertake the research required to acquire this understanding, when no single party had central oversight and control, and to update understanding as new information emerged.

6.2 Intended use cases

Key to the appropriate evaluation of tests to inform testing policy was the specification of intended use cases (see Section 5.1). In the Covid-19 pandemic, it was important to differentiate testing:

- Of symptomatic from apparently healthy people.
- By where tests are used—in community, primary care or secondary care settings.
- When tests are used in relation to onset of symptoms.
- By whether strategies include repeat or confirmatory tests.
- By the nature of the biological sample.
- By the process for collecting specimens.
- By whomsoever processes the tests, particularly if tests are for self-use.
- By the target condition being diagnosed.
- By the scale and timescale of testing.
- And by whether results are to be used to determine patient management, control disease, and/or for surveillance purposes.

Several organizations (e.g. FIND, 2021) created descriptions of intended use cases for Covid-19 tests. Ideally, each intended use case for a test requires its own evaluation, undertaken in the appropriately matched real-world setting.

Whereas extrapolation of results from one use case to another might sometimes be considered, the performance of tests can vary between use cases (Example 1, Section 5.2.1) emphasizing the importance of undertaking evaluations of tests for each intended use; and monitoring their performance during implementation. Any such extrapolation should be explicit about the assumptions made.

During the pandemic, new potential intended use cases emerged as new considerations were made about how best to tackle disease spread, whereas others—originally thought to be important—were found to be unnecessary, impractical or impossible (Example 2, Section 5.2.3).

Example 1 (Extrapolation of performance of antigen tests from symptomatic to mass testing). The initial Innova lateral flow test evaluations reported by the University of Oxford and Public Health England were undertaken in regional test-and-trace centres where recruited participants were expected to have symptoms. Estimated sensitivity compared to the reverse transcription PCR (RT-PCR) reference standard was between 58% (95% CI: 52 to 63) when tests were run by test-and-trace centre staff and 79% (95% CI: 72 to 84) when tests were run by laboratory scientists (Peto et al., 2021). A decision was then made to pilot use of the Innova test for mass screening of people without symptoms (University of Liverpool, 2020).

A subsequent Cochrane review showed that the sensitivity of other lateral flow tests (LFT) to detect SARS-CoV-2 in people without symptoms had, on average, sensitivity between 15% and 20% points lower than in people with symptoms (Dinnes et al., 2021). For example, based on testing at Wisconsin University Campus, the Sofia antigen test had a sensitivity of 80% (95% CI: 64 to 91) when used in 227 people with symptoms, and only 41% (95% CI: 18 to 67) when used in 871 people without symptoms (Pray et al., 2021).

In the UK, the mass screening of citizens without symptoms in Liverpool with the Innova test included a dual swabbing evaluation (LFT and PCR), for which around 6,000 citizens gave consent and which also reported a lower sensitivity of 40% (95% CI: 29 to 52) (University of Liverpool, 2020). The higher accuracy in people with symptoms likely relates to testing occurring whilst viral levels are close to their peak soon after the onset of symptoms. See further detail in Example 8.

Example 2 (Initial roles for antibody tests were abandoned). Early in the pandemic, the UK focused on procuring point-of-care antibody tests, and commissioned evaluations of tests imported from China as it could be important to identify individuals with antibodies who might be immune (Boseley, 2020; Royal Society of Medicine, 2020).

Initial point-of-care antibody test performance was deemed inadequate (Adams et al., 2020) and so, instead, the Government purchased substantial quantities of laboratory-based antibody tests which have formed the NHS Pillar III testing programme. However, concerns emerged that antibody levels might not endure so that, except in surveillance studies, there was little point in individuals being tested.

6.2.1 Intended use cases for molecular and antigen tests

Molecular (e.g. PCR) and antigen tests aim to identify who is infected with SARS-CoV-2. The clinical use case for these tests was in

(a) Symptomatic people to diagnose Covid-19.

However, they also had public health use cases for:

(b) Testing contacts of cases.
(c) Identifying outbreaks.
(d) Screening apparently healthy individuals to find asymptomatic cases who might nevertheless transmit infection.
(e) Ruling out current infection.
(f) In surveillance studies that estimate the prevalence of infection.

Thus, the target condition was either SARS-CoV-2 infection (if asymptomatic) or Covid-19 (if symptomatic). In addition, mass testing was proposed to identify individuals who were infectious rather than just infected. Problems in considering infectiousness as a target condition are discussed

in Sections 6.6.2 and 5.3.3. The above list of use cases again relates to different participant groups: either symptomatic or asymptomatic, with or without known exposure, and to the timing of tests.

There were several different types of tests which varied in their performance, the laboratory facilities and staff required to deliver the test, potential testing capacity, the cost and accessibility of the test, and the time taken between obtaining the specimen and the result. Most tests required a throat and/or nasal swab; some were trialled on saliva samples. Decisions about the use of different tests thus depended on performance but also on cost, speed, capacity and accessibility.

6.2.2 Intended use cases for antibody tests

Antibody tests identify whether an individual has developed antibodies to SARS-CoV-2. These tests were considered for clinical and public health purposes. Clinical purposes include:

(a) Identifying SARS-CoV-2 in people with prolonged symptoms of Covid-19 who have presented too late for an antigen test to be able to detect the virus.
(b) Assessing whether individuals are mounting an antibody response, either to the disease or a vaccine.
(c) Assessing whether individuals have sufficient levels of antibodies to be considered as a plasma donor.

Each of these cases relates to a different group: (a) people with symptoms but it is not known if they have Covid-19; (b) people known to have Covid-19, or recently vaccinated; (c) people who have recovered from Covid-19. For public health and research purposes, antibody tests can be used to estimate:

(d) Within surveillance studies, how many have previously had disease.
(e) The persistence of antibodies.

These again require different participant groups: (d) unselected population samples; and (e) those known to have been infected.

6.3 Study designs

Section 4.1 outlined the limitations of analytical studies which assess the properties of a test on samples in laboratory settings, and Section 4.2.1 described studies which pre-select participants who are already known to have or not to have the target condition. Although studies of analytical validity allow speedy assessment of the potential diagnostic performance of a test, they do not provide evidence of its performance in a real-world setting.

6.3.1 Regulatory requirements for evidence have varied

New tests were developed at pace to address the emergency need for diagnostics during the Covid-19 pandemic. The standard process described in Section 4.1 of undertaking laboratory studies of the key analytical properties of new tests prior to assessing their accuracy in real-world field settings was followed. However, evidence from analytical studies was the main evidence considered for many applications for Emergency Use Authorization (EUA) marketing approval. Evaluations of clinical performance in real world studies of tests as used for their intended uses often followed later, sometimes alongside their implementation, but sometimes not at all.

Evidence requirements are not consistent across regulators, and changed during the pandemic (Shuren & Stenzel, 2020). The Conformité Européenne (CE) IVD marking used across the UK and European Union (EU) is primarily a 'declaration of conformity' with EU requirements, and not an application which undergoes scrutiny (Allan et al., 2018). Although the IVD Directive 98–79 (European Parliament & Council, 1998) mentions both analytical and diagnostic sensitivity and specificity, it does not define these, nor require a 'use case' to be stated. Establishing analytical performance has appeared adequate to obtain the CE-IVD mark necessary to enter the market, without the need for field studies evaluating performance for established intended use cases. The Food and Drugs Administration (FDA) has a more rigorous process but has frequently allowed tests to market based on EUAs reliant on similar analytical performance evidence (see

Example 3, Section 4.1). For a period during 2020, all antibody tests were allowed to be marketed in the United States of America (USA) without restriction, a decision which allowed poor-performing tests onto the market (Shuren & Stenzel, 2021).

Example 3 (Test implementation based on analytical performance). Initial approval of the Abbott ID-NOW test for current infection by the FDA was based on analysis using spiked samples without any evaluation in humans. Data in the IFU showed that the test had 100% (83.9 to 100) sensitivity in 20 samples at viral concentrations twice the limit of detection, and 100% (88.7 to 100) specificity in 30 samples with no virus (Abbott Diagnostics, 2020). The subsequent Cochrane review (Dinnes et al., 2021) reported the Abbott ID-NOW test sensitivity in real-world use to be 73% (69 to 78) with specificity of 99.7% (98.7 to 99.9) based on results from four studies with 812 samples including 222 SARS-CoV-2 cases.

Reportedly, the Abbott ID-NOW was used in Washington to test attendees at the White House Rose Garden on the 26th September 2020 (Mandavilli, 2020), after which at least 31 people were infected including President Trump (who was hospitalized for 3 days) and the First Lady (Buchanan et al., 2020).

6.3.2 *Analytical and two-group study estimates of sensitivity and specificity*

Manufacturers usually report estimates of test sensitivity and specificity in their IFU documents and separate claims of analytical performance from those of clinical performance. However, this distinction is often not made clear to the public on websites and in advertising, and rarely are studies reported that directly fit with an intended use for the test. This reporting failure is despite the 2019 International Organization for Standardization (ISO) statement on studies of clinical performance (ISO, 2019) which emphasizes the importance of stating the intended use of an in vitro medical device and proving, in that context, how the test relates to a particular clinical condition or physiological/pathological process/state. Manufacturers' websites and IFUs also rarely provide adequate details to ascertain the source and characteristics of the participants in evaluation studies (see Example 4), or to clarify whether individual participants provided single or multiple specimens. Rarely is it clear whether estimates of analytical sensitivity and specificity are based on laboratory samples, whether estimates of diagnostic sensitivity are from two-group studies using preselected specimen banks of known disease status or collected prospectively for particular use cases (Bigio et al., 2023) (see Example 4, Section 4.2.1).

Example 4 (Marketing claims based on selected samples). The UK Rapid Test Consortium AbC-19 rapid antibody test was initially marketed with claims of sensitivity of 98.03% (95.03 to 99.46) and specificity of 99.56% (98.40 to 99.95) on the manufacturer's website (Abingdon Health, 2021a). A subsequent preprint showed that the samples were sourced from a mixture of biobanks and cohorts, and that the sensitivity was based on 'known positive' samples pre-selected as positive on two of three other antibody tests; specificity was evaluated on 'known negatives' pre-selected as negative on all three other antibody tests; all other samples were inadmissible (Robertson et al., 2021). Selective sampling, wherein samples most likely to show positivity or negativity are purposely chosen (and the more difficult to detect or rule-out cases are omitted), leads to bias (see 4.2.1). See also Example 9.

Later nonselective evaluation by Public Health England found lower sensitivity of 81.5% (77 to 85) and specificity of 99% (98.5 to 99.4) in a cohort study using the same laboratory-based immunoassays as the reference standard but including all participants (Mulchandani et al., 2020).

6.4 Participants

Studies of SARS-CoV-2 tests have shown how the performance of tests depends on the population subgroups to whom they are applied. The sensitivity of the test is determined by the characteristics

of individuals with the condition and the specificity of those without the condition. Given the pattern of viral and antibody kinetics, it can be foreseen how testing at different time points will affect test performance: testing before or after the peak in viral load will increase the risks of FNs in antigen tests which cannot detect lower levels of the virus; testing before antibodies rise will increase FNs for antibody tests; and testing when the prevalence of infection is low will increase the proportion of test positives that are FPs.

For example, antigen tests are known to miss cases when viral loads are low; thus, their sensitivity will be lower if used in groups and at time points when viral loads are lower or high for only short periods of time. Differences noted between the performance of antigen tests in symptomatic vs. asymptomatic groups (see Example 1, Section 5.2.1) (Dinnes et al., 2021) are potentially explained by evidence that peak viral loads are of shorter duration in those who are asymptomatic (Cevik et al., 2021).

Similarly, antibody tests miss cases with no or low antibody responses and were less accurate when used soon after symptom onset than later (Example 5). If antibody response relates to disease severity, the test's performance will differ between those with a disease severe enough to warrant hospitalization and those who stay at home or are asymptomatic (Post et al., 2020).

Compared to evaluation in groups with no known condition, evaluating the specificity of the test in symptomatic groups, which include individuals with diseases caused by similar respiratory-based viruses, may increase the risk of producing FP results if the test fails to distinguish between similar viruses. In the first Cochrane review of antibody tests, the FP rate for IgG tests in six studies ($n = 396$) that recruited individuals suspected of having Covid-19 was 2.0% (95% confidence interval (CI): 0.4 to 9.0) compared to 0.8% (0.2 to 2.4) in 10 studies ($n = 2{,}614$) that included currently healthy participants, and 0.8% (0.3 to 2.2) in 10 studies ($n = 2{,}633$) on pre-pandemic cohorts (Deeks et al., 2020). Notice, however, that the confidence intervals here are too wide for inferences to be drawn.

6.5 Index test

6.5.1 Timing of testing

The performance of antibody tests relies on their use after antibodies initially rise, and before they wane. In the early period of the pandemic, the use of antibody tests was considered as a diagnostic test in those presenting with symptoms; but had very poor sensitivity for that context of use (Example 2, Section 5.2.3). When tests were used at a later time point, higher levels of sensitivity were obtained (see Example 5). Some antibody tests may have been inappropriately abandoned due to their lack of accuracy in the early time period.

As the pandemic progressed, the waning of antibody responses will have affected the ability of surveillance studies based on antibody tests alone to identify previously infected cases, as they would only have identified those in whom antibodies were still detectable. Evidence has accumulated on the likely duration and variability in antibody response between tests and between individuals (Ward et al., 2021a). Equally, it has become difficult to identify those with responses to infection from those with responses to vaccination (Ward et al., 2021b).

6.5.2 Test samples, methods and operators

Variations in the samples used, the process by which they are obtained, and the execution of tests can affect performance. For example, the sensitivity of the Oncogene reverse transcription loop-mediated isothermal amplification (RT-LAMP) test which has been compared on real and spiked samples (Example 6, Section 4.1), and on swab and saliva samples, with and without ribonucleic acid (RNA) extraction stages, was found to vary between 70% and 95% (Example 7, Section 5.2.4) (Department of Health & Social Care, 2020a).

There are also many studies which have used specimen types without the manufacturer's approval (such as saliva or viral transport media (VTM) when dry swabs are required), or blood samples taken in different ways (venous vs. skin prick for antibody tests—see Example 8). Variations have also been noted in the performance of tests according to the level of expertise of the tester, for example where tests involve multiple steps and/or a degree of subjective interpretation.

The variation in the Innova test's reported sensitivity illustrates the potential for combinations of population, tested material, and testing operative to impact the sensitivity of an antigen test (Example 9, Sections 4.2.1, 5.2.1, 5.2.4).

Example 5 (Impact of time since symptom onset on positivity of antibody tests). Antibody tests have proven easier to manufacture than antigen tests and were initially evaluated to see whether they could be used on presentation of symptomatic cases at Accident and Emergency (A&E) Departments and other health facilities for initial diagnosis of Covid-19. Cassaniti et al. (2020) used an IgG test in 50 individuals presenting with fever and respiratory symptoms indicative of Covid-19 at an Italian hospital's A&E. Only 18% (95% CI: 8 to 34) of those found to be positive on PCR for SARS-CoV-2 were antibody positive on the test, most likely because the test was being used before antibody levels were detectable. In a second part of the study, the same test was used in 30 different individuals admitted to Intensive Care where 83% (95% CI: 65 to 94) were positive when tested a median (interquartile range) of 7 (4 to 11) days after their first test.

The Cochrane review (Deeks et al., 2020) of antibody tests for SARS-CoV-2 included an analysis which showed a strong time-trend of increasing sensitivity with time since symptoms: up to 90% from 14 days post symptom onset (see Table below). Thus, antibody tests may have a diagnostic role in recognizing Covid-19 in people presenting very late after onset of symptoms to be detected by an antigen test, but not earlier than 10–14 days.

	Sensitivity (95% CI)				
	Days since onset of symptoms				
	Days 1–7	Days 8–14	Days 15–21	Days 22–35	Days > 35
IgG	30% (20 to 39)	67% (58 to 74)	88% (84 to 92)	80% (72 to 86)	87% (80 to 92)
	23 studies	22 studies	22 studies	12 studies	4 studies
	568 samples	1,200 samples	1,110 samples	502 samples	252 samples
IgG/IgM	30% (21 to 41)	72% (64 to 80)	91% (87 to 94)	96% (91 to 98)	78% (66 to 86)
	9 studies	9 studies	9 studies	5 studies	2 studies
	259 samples	608 samples	692 samples	52 samples	53 samples

Example 6 (Differences between real and spiked samples). The Department of Health and Social Care (DHSC) evaluated the RT-LAMP test (see also Example 7) on saliva and swab samples (Department of Health & Social Care, 2020a). Due to difficulties in obtaining adequate saliva samples from Covid-19 positive individuals ($n = 167$), the number of samples was increased by addition of samples spiked with SARS-CoV-2 ($n = 59$) in the laboratory. The overall estimate of sensitivity combined both real and spiked samples ($n = 226$).

Sensitivity in real vs. spiked samples, when stratified by viral level (as defined by the cycle threshold (Ct) value from the accompanying RT-PCR test), showed disparities. At the lowest viral loads (highest Ct values), the RT-LAMP saliva test detected 13% (95% CI: 16 to 38) in real samples compared to 91% (78 to 97) in spiked samples. Estimates based on spiked samples cannot be considered as estimating how the test will perform in real world settings.

Viral load	Sensitivity (95% CI) real cases	Sensitivity (95% CI) spiked samples
Ct < 25	Cases (74/79): 94% (86% to 98%)	Spiked (9/9): 100% (66% to 100%)
25 ≤ Ct < 33	Cases (51/72): 71% (59% to 81%)	Spiked (3/6): 50% (12% to 88%)
33 ≤ Ct < 45	Cases (2/16): 13% (16% to 38%)	Spiked (40/44): 91% (78% to 97%)

(Sensitivity estimates calculated from data in the report and Supplementary Tables).

Example 7 (Impact of sample type and processing on test sensitivity). RT-LAMP tests are potential alternatives to RT-PCR tests for detecting SARS-CoV-2 and utilize a faster isothermal process. Two adaptations which could further increase the usability of RT-LAMP are testing of saliva rather than a nasopharyngeal swab, and direct testing of the sample rather than testing RNA extracted from the sample.

The DHSC reported a study of the Oncogene RT-LAMP tests comparing performance on swab samples (nasopharyngeal swabs/oropharyngeal swabs) with saliva samples, and comparing testing extracted RNA samples and crude clinical samples (Department of Health & Social Care, 2020a). (It is unclear how the samples were allocated to different testing methods, or whether they were different or the same samples). The four combinations produced the following results:

Sample type	True Positive (TP)/Covid-19 cases	Sensitivity (95% confidence interval)
RNA on swabs	179/188	Sensitivity of **95%**: 95% CI (91% to 98%)
RNA on saliva	89/111	Sensitivity of **80%**: 95% CI (72% to 87%)
Direct swabs	140/199	Sensitivity of **70%**: 95% CI (63% to 77%)
Direct saliva	127/167	Sensitivity of **76%**: 95% CI (69% to 82%)[a]

[a]Calculated excluding the 59 spiked samples mentioned in Example 6.

The risk that a test will miss an infection that is present is computed as 1-sensitivity.

For testing of swab samples, the risk increased from 5% for RNA from swabs to 30% by omitting the RNA extraction step, an increase of 25% points (95% CI: 18% to 32%). FNs also increased by 15% points (95% CI: 7% to 23%) when using RNA from saliva samples instead of RNA extracted from swabs.

6.5.3 Appropriate samples and settings

Both laboratory based and point-of-care lateral flow antibody tests have been developed during the pandemic. Laboratory based tests have used enzyme-linked immunosorbent assay (ELISA) or chemiluminescence enzyme immunoassay (CLIA) based techniques, often designed to run on large analytical platforms that are already installed in clinical laboratories and are capable of running multiple tests at the same time. Laboratory based tests have been developed to utilize venous blood samples. Point-of-care lateral flow antibody tests have been developed to run on capillary

finger-prick blood samples but have not received regulatory approval for home use. There have been issues in ensuring that tests are sold and used on the samples for which their use has been evaluated (Example 8, Section 5.2.4).

Example 8 (Importance of approved and evaluated use for sample types). Commercial suppliers in the UK were keen to sell antibody test services direct to the public but, as there were no lateral flow assays licensed for home use, they decided to collect capillary finger prick blood (which individuals can obtain themselves) to run on laboratory machines. Sales were suspended by the regulator as capillary blood was not an approved specimen-type for the laboratory machines—at least until further evaluation studies were undertaken to establish the performance of the test on finger-prick blood (Medicines & Healthcare products Regulatory Agency, 2020a).

The performance of POCTs differed when evaluated in laboratory settings on serum from venous blood samples compared to real world settings using the intended self-read finger prick samples. For example, using serum and finger-prick samples from the same people, the AbC-19 antibody test was positive in 46 out of 50 (92%; 95% CI 81 to 98) serum samples in the laboratory, but in only 32 of 51 (63%; 95% CI 48 to 76) self-read finger-prick samples in the clinic (Moshe et al., 2021).

6.5.4 Repeated testing—importance of correlation of test errors

Rapid antigen test strategies for use in asymptomatic persons may involve repeatedly testing individuals at varying frequencies, including daily use. The accuracy of the strategy of repeated tests required specific evaluation: only one field study of daily antigen testing had been reported by March 2021 (by which time they were being used in schools) (Dinnes et al., 2021). Many estimates of predicted performances are based on naïve extrapolation from a single test used under an independence assumption, with policy being based upon mathematical models rather than empirical evaluations.

Naïve Bayes estimation of serial test performance assumes independence between successive tests. Assuming independence allows the multiplication of the probabilities of FNs (or FPs) as a means for estimating serial performance. For example, if independence held and infection-status does not alter, applying test X with a FN rate of 30% twice would yield an overall 9% FN rate (0.3×0.3) from two applications of test X, or 2.7% if the test was used three times (Ramdas et al., 2020).

Independence assumes that the performance of test X at each time point in each individual is unrelated to its performance at previous time-points in the same individual (Deeks et al., 2021b). However, as the FN rate for SARS-CoV-2 antigen test X is likely to relate to a person's viral trajectory, correlation of test results between close time-points is expected within-individuals and independence unlikely. Consider two individuals who have become infected and were tested everyday with RT-PCR. Observed Ct values (higher values = lower viral load, see Section 6.6.3 for a fuller description) for days 1–7 for individual A were 35, 28, 23, 12, 12, 15, 20 and for individual B were 38, 35, 35, 37, 28, 26, 28. Individual A reported a higher viral load quickly which was maintained, whereas individual B had a longer latent period, and a lower peak viral load. If these individuals were tested using an antigen test which (for the sake of simplicity but without compromising the message) could only detect the virus when viral loads were such that Ct values were <25, the test results over seven days would be FN, FN, TP, TP, TP, TP, TP for A and FN, FN, FN, FN, FN, FN, FN for B. It is quite clear that we should expect to see 'runs' of either FNs or TPs and not random sequences, so that the probability of obtaining a TP or FN at any time-point relates to results at previous time-points, particularly those that are close.

Simulating realistic data which match the evolving correlations between time-points is challenging. For example, some agent-based models have addressed this by simulating underlying viral load trajectories (Larremore et al., 2021; Quilty et al., 2021). It is important that robust randomized empirical evaluations of serial-testing strategies are undertaken as any modelling of the underlying biology will necessarily rely on simplifying assumptions.

Example 9 (Differences in estimates of sensitivity: Innova test). There were assessments of the sensitivity of the Innova Lateral Flow Rapid Antigen test in different patient groups, using samples stored and processed in different ways, and delivered by differently trained testers/readers. Studies showed variation in sensitivity from 96% (when used in patients admitted to hospital with pneumonia within 5 days of symptom onset) to 3% (when used to screen asymptomatic university students).

Results in asymptomatic groups (Liverpool, University of Birmingham) had lower sensitivity than those in symptomatic groups. There were also differences between testing done on fresh swab samples vs. testing on frozen samples or by use of the viral transport media (VTM); and whether the tests were run/read by laboratory professionals, healthcare workers (HCWs) or trained non health care workers.

Study	Participants	Setting	Sample	Tester	TP/ Covid-19 cases	Sensitivity (95% CI)
IFU [1]	Pneumonia (<5 days symptoms)	Inpatients	Dry swab[a]	Not stated	72/75	96% (89 to 99)
PHE [2]	SARS-CoV-2 positive patients	Hospitalized patients	Frozen VTM fluid in saliva	Lab	95/178	53% (46 to 61)
PHE Falcon [2]	Symptomatic	Test-and-Trace centre	VTM fluid	Lab	156/198	79% (72 to 84)
PHE Falcon [2]	Symptomatic	Test-and-Trace centre	Dry swab[a]	HCW	156/223	70% (63 to 76)
PHE Phase 4 [2]	Symptomatic	Test-and-Trace centre	Dry swab[a]	non-HCW	214/372	58% (52 to 63)
PHE Phase 4 [2,3]	Not stated	Navy barrack outbreak	VTM fluid	Lab	13/46	28% (16 to 43)
Liverpool [4]	Asymptomatic	Mass testing	Dry swab[a]	non-HCW	28/70	40% (28 to 52)
Uni B'ham [5]	Asymptomatic	Student testing	Dry swab[a]	non-HCW	2/62[a]	3%[b] (1 to 16)

[a]Tested according to manufacturer's instructions.
[b]This study sampled 10% of noncases. Total Covid-19 numbers, sensitivity and its confidence interval are computed by reweighting for the study design (see Example 11).

IFU = Instructions for Use; PHE = Public Health England; Uni B'ham = University of Birmingham; Lab = tested by scientists in laboratory at Porton Down; HCW = tested by trained health care workers; non-HCW = tested by trained staff working at testing centre; VTM = viral transport medium.

References: [1] Innova Medical Group (2020); [2] Peto et al. (2021); [3] Dinnes et al. (2021); [4] University of Liverpool (2020); [5] Ferguson et al. (2021).

6.6 Target conditions and reference standards

As viruses mutate and produce new variants, the sensitivity of existing tests may change. Whilst molecular understanding of mutations may inform the impact on performance, empirical verification is required. In many instances, laboratory studies may be adequate to confirm sensitivity, but

if a new variant does have an impact in the laboratory context, then field evaluations of the tests or modified tests are likely to be required to provide confident estimates of test performance.

6.6.1 *Different target conditions for different use cases*

Although rarely made explicit, statements about the performance of a test are all pertinent to a particular target condition (assessed using a particular reference standard), and it is to be expected that the performance of a test will not be the same for different target conditions (Lord et al., 2011).

For the use cases (a) to (e) for antibody tests (see Section 6.2.1), there are three different target conditions: (a) considers whether individuals are currently infected; (d) whether they were infected previously; whereas for (b), (c), and (e) the target condition is the presence antibodies (see Example 10, Section 5.1).

Reference standards are the best method for identifying whether individuals do or do not have the target condition. For current or previous infection, the reference standard should relate to whether an individual has currently, or has a history of, proven SARS-CoV-2 infection, typically evidenced by one or more RT-PCR tests or fulfilling the case definition for SARS-CoV-2 (World Health Organization, 2020).

Equal attention needs to be paid to ensure the reference standard classification of those never infected is accurate, either by multiple negative RT-PCR tests, or clear history that no infection could have been possible (often achieved using prepandemic sera banks). When tests are used for surveillance, statistical methods to correct for misclassification rates should be applied to obtain accurate population estimates (Diggle, 2011). Antigen tests which themselves have high performance may be used to assess the performance of tests for the target condition of presence of current antibodies.

Example 10 (Mismatch in target condition between evaluation and intended use). There are examples of mismatches between reference standards, target conditions and intended use. For example, the UK Government purchased the AbC-19 antibody test from the UK Rapid Test Consortium to be used for 'surveillance studies to help build a picture of how the virus has spread across the country' (Department of Health & Social Care, 2020b). The implied target condition to be assessed is previous infection. However, the manufacturer states that the test 'is not designed to detect previous infection but rather to detect the presence of a particular type of antibody' (Abingdon Health, 2021b). The manufacturer assessed the test in a study of known antibody positive and known antibody negative samples (see Example 4). Individuals who had previous RT-PCR confirmed infection but developed no or very low antibody levels were excluded from the disease positive group [14 of 265 (5%)] (Robertson et al., 2021). Thus, the manufacturers' estimates of sensitivity and specificity relate to the ability to detect antibodies; and not for a surveillance role to detect previous infection.

6.6.2 *Infectiousness—a different target condition without a reference standard*

Management of transmission requires isolation of people who are or will become infectious to prevent their transmitting the virus to others. Infected individuals may be infectious for a number of days, and some were identified as more infectious than others (i.e. 'super spreaders'). Identifying how accurately a test identifies those who are infectious requires undertaking studies where infectiousness is the target condition, assessed using a reference standard which accurately classifies people according to whether they could or could not transmit the virus to others. Claims were made that rapid antigen tests identify individuals who are infectious or are most likely to be infectious.

However, there is no reference standard for infectiousness (see Section 5.3.1). Direct evidence of transmission to secondary cases clearly indicates infectiousness, but absence of transmission does not indicate noninfectiousness, particularly when people have been isolating (and thus preventing transmission which otherwise would have occurred).

It was also argued that infectious people must have viable virus (in that it can replicate). Viral viability was thus assessed in patient samples by attempting isolation in cell culture, but viral culture is difficult to run on account of biohazards and necessarily high levels of laboratory precautions. Viral culture is also known not to be sensitive; and is dependent on operator and laboratory expertise [studies of SARS-Cov-2 culture in different laboratories have shown variation in success rate, e.g. Bullard et al (2021) found 26/90 (29%; 95% CI 20 to 39) whereas Singanayagam et al (2020) found 134/324 (41%; 95% CI 36 to 47)].

Ct values from RT-PCR relate to viral load. As rapid antigen tests only detect those with higher viral load, studies have attempted to establish the relationship between Ct value and markers of infectiousness (both secondary case rates and viral culture), and to link the performance of rapid antigen tests to Ct level to see whether they could accurately detect infectious people (Deeks et al., 2022). There are, however, very few studies which have attempted viral culture in individuals receiving rapid antigen tests, (e.g. Schuit et al., 2021). Put simply, maximum Ct values (i.e. minimum viral loads) identified in the studies above in which secondary cases or viral culture do not occur were denoted as 'infectiousness thresholds' to classify individuals as having positive or negative 'infectiousness status'. Three statistical issues arise: the first is disregard for the statistical challenges in estimating extreme values (Haan & Ferreira, 2007); the second is that 'infectiousness threshold' needs external validation before being adopted more widely; and third is that the empirical data display a continuity of decreasing risk with increasing Cts without any clear lower bound (Singanayagam et al., 2020—also see Example 12). However, policy appears to have been constructed from stating binary thresholds. For example, Scientific Advisory Group for Emergencies (SAGE) minutes report 'expert opinion … suggests that a Ct value of below 25 seems to be associated with viable transmission'(Scientific Advisory Group for Emergencies, 2020), and the website of Innova stated 'According to published results from University of Oxford and Public Health England (PHE) clinical study, Innova's rapid antigen tests have roughly a 97% efficacy in detecting infectious patients' based on interpreting data below a Ct of 25 as infectious (Innova Medical Group, 2021).

The regulator stated that tests in asymptomatic individuals need to be assessed for the target condition of current infection defined as an infection in which the causative organism is live and has the potential, either now or in the future, to cause disease or onward transmission (Medicines & Healthcare products Regulatory Agency, 2020b). This is more inclusive than infectiousness, but equally there was no clear reference standard at the time that could clearly differentiate between current and recent or previous infection.

6.6.3 Establishing the performance of RT-PCR

The RT-PCR test was the primary diagnostic test for SARS-CoV-2 infection used globally from the start of the pandemic and became established as the reference standard against which other tests were compared. However, questions were raised concerning the performance of the RT-PCR test, particularly its FP rate. We note that the test evaluation paradigm does not allow easily for estimation of the accuracy of a test considered as the reference standard. Alternative analytical approaches are required: for example, an upper bound can be placed on the potential FP rate with RT-PCR by considering the total positive rate when disease prevalence is low (see Example 11).

Example 11 (Estimating the performance of RT-PCR without a reference standard). UK population prevalence surveys using RT-PCR have shown test positivity rates from the Office of National Statistics (ONS) Community Infection Survey in England of 0.44% (95% credible interval: 0.22 to 0.76) in August 2020 (Office for National Statistics, 2020), and 0.077% (95% confidence interval: 0.065 to 0.092) from the Real-time Assessment of Community Transmission (REACT-1) study in June to July 2020 (REACT et al., 2020). Lower rates have been observed in other countries such as Australia (Our World in Data, 2021) and China (Cao et al., 2020). These figures place an upper bound on the specificity of the test.

The NPV of RT-PCR was estimated in studies which retest symptomatic people with a second RT-PCR who initially tested negative on their first RT-PCR. This does not directly assess sensitivity; but explores the impact

of FNs. A review of such studies showed variation between 2% and 54% of negative PCR tests being FNs, with a pooled estimate of 13% (95% CI 9% to 19%) (Arevalo-Rodriguez et al., 2020).

The RT-PCR test is quantitative, in that it produces a count of the number of amplification cycles that the process has moved through before a sample showed a response (Ct value). As discussed above, this Ct value relates to the viral load for that biological sample, low numbers indicating higher viral loads. Comparing test performance stratified by Ct value became a key issue for antigen tests but was hampered by an absence of standardization of Ct values between machines. Studies to understand the relationship between Ct values and viral load were based either on using samples with varying known serial dilutions of a SARS-CoV-2 plasmid viral load, measuring the Ct value and constructing a fitted curve; or using quantitative viral load methods on patient samples with known Ct values (Case et al., 2020) (see Example 12). The relationships were then estimated using linear regression, some regressing $\log_{10}$ viral load on Ct value (e.g. Mitja et al., 2021), others Ct value on $\log_{10}$ viral load. Rarely was detail given about the uncertainty in estimates, the suitability of a linear relationship, or the statistical fit of models—all of which are remediable problems. Little attention was paid to use of standard measurement comparison methods (such as Bland–Altman methods, see Section 4.1.2) to estimate systematic differences in Ct values between different machines. Nor were there many studies of measurement error in Ct values.

Example 12 (Relationship between viral load and viral culture). PHE scientists attempted virus culture on 324 upper respiratory track samples from 253 people who were suspected of having Covid-19 and in whom the virus was detected on RT-PCR (Singanayagam et al., 2020). Samples were obtained from a range of clinical scenarios including community and healthcare worker surveillance, symptomatic persons tested as part of the early epidemic response and samples acquired in outbreak investigations. Vero E6 cells *(a standard cell line used for growing viruses in cell culture)* were inoculated with clinical specimens and incubated and inspected for cytopathic effect daily up to 14 days. Viable virus was isolated from 134 (41%) samples (from 111 cases). Considering the culture positivity rates against the Ct values from PCR showed the expected decline in culture rates with increasing Ct value (indicating decreasing viral load), but culture was achieved down to Ct levels of 37, well below the detection level of rapid antigen tests (typically between 10^5 and 10^7 viral particles per ml, equivalent to Ct values of 20–27) (see Figure 5). At low viral loads, corresponding to Ct values above the oft quoted Ct = 25 threshold, virus could still be cultured from one-third (93 of 276) of samples. Other studies, in larger samples, showed similar relationships. By fitting a model, rather than estimating minimal values or percentiles, the authors provide data that allows probabilistic assessment that the virus can be cultured from samples at each Ct value.

6.7 Explaining summary statistics for each use case

Whilst the performance of each test is typically summarized as sensitivity and specificity (which are probabilities conditional on infection status), the individuals who are tested also need to understand the meaning and implications of the positive and negative test results that they receive, which are described by probabilities computed conditional on test result (the predictive values, Section 4.2.3). Public health discussion often naturally focused on the sensitivity as it describes the proportion of cases that will be detected, and the specificity as it describes the risk of FPs, but it is equally important to ensure that the public, press and policymakers understand, at an individual level, the implications of positive and negative results. Population benefit of testing can only be fully realised when the responses of individuals receiving test results are appropriate.

When explaining probabilities to the public it is essential that the two different types of error (FNs and FPs) are explained clearly to avoid confusion. Good practice is to ensure that the probabilities related to FPs and FNs are explicitly explained, rather than relying on the public's understanding of sometimes arcane statistical terminology (see Example 13). For example, the phrase

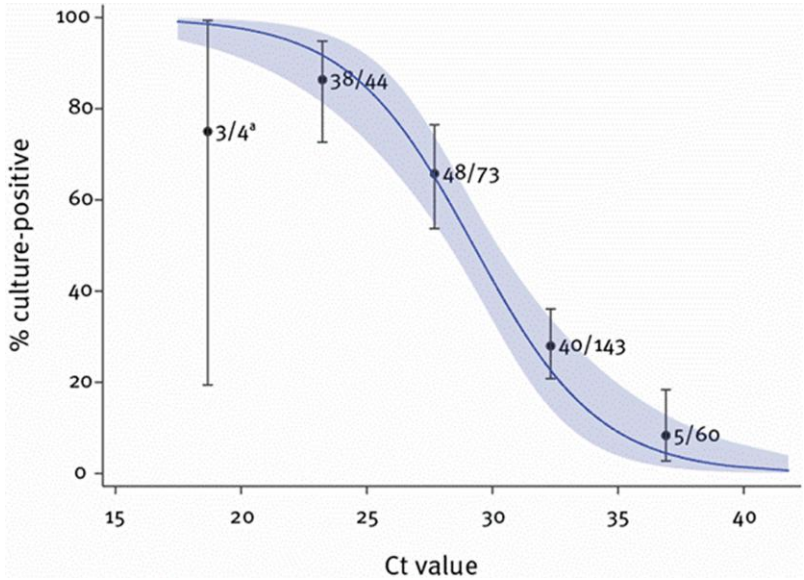


Figure 5. SARS-CoV-2 virus culture positivity rates from upper respiratory tract samples from 324 people suspected of having COVID-19, grouped by Ct values from RT-PCR (excerpt from Singanayagam et al., 2020; distributed under creative commons attribution license). Bars represent 95% confidence intervals.

'false positive rate' is sometimes used for both for 1-specificity and 1-PPV, which can cause confusion. When infection events are rare (i.e. when prevalence is low), these two probabilities differ considerably and, if confused, give a seriously wrong impression. When the prior probability of infection is very low, even tests which have exceptionally high specificity can give more FPs than TPs (see Example 14, Section 4.2.3).

Example 13 (Poor communication about test performance sent to schools). The Department for Education guide to schools (NHS Test & Trace, December, 15th 2020) summarized the performance of the Innova lateral flow test by stating 'These tests work … they were shown to be as accurate in identifying a case as a PCR test (99.8% specificity). The tests have lower sensitivity but they are better at picking up cases when a person has higher viral load'.

This statement by Test-and-Trace and the Department for Education did not withstand statistical scrutiny (Deeks et al., 2021a) and was later removed. In particular, the word 'accurate' was likely to be interpreted by the public as meaning 'few errors of any type'. It is not clear whether the ability to 'identify a case' refers to the ability to detect infections which are present (test sensitivity) or to positive results implying you have an infection (PPV). Neither was described by test specificity, which was the only probability quoted.

It is also important that the diagnostic value of negative test results is properly explained. When the disease is rare, NPVs can give a misleading impression as the probabilities are all close to one. For example, using the estimates of the accuracy of Innova from Liverpool (sensitivity of 40% and specificity of 99.9%), if the prevalence of the disease is 1 in 1,000, those testing negative on Innova have a post-test probability of 99.94% of not having an infection, suggesting that a negative test if very helpful. However, this must be compared with the chances of infection in those untested of 99.9% to reveal how little the chance of infection has been reduced by getting a negative test result.

Simple computation of likelihood ratios (LRs) shows that whilst a positive test result indicates that the relative chance of disease has greatly increased (LR+ = sensitivity/(1 − specificity) = 0.4/0.001 = 400), a negative test-result makes little relative difference (LR− = (1 − sensitivity)/

specificity = 0.6/0.999 = 0.6). This is one occasion where a likelihood ratio may be a good way of explaining the value of a test result: 'getting a negative result does not even halve the prior probability that you have the infection' (as events are rare, it is not necessary to use odds in this expression as they approximate closely to probabilities).

Example 14 (Impact of prevalence on the need for confirmatory testing). Monitored asymptomatic screening (thrice in 2 weeks) of pupils on their return to secondary schools in England in early March 2021 used antigen lateral flow tests—but was introduced without confirmatory testing of positives by RT-PCR. Individuals who tested positive, together with their family and contacts in their school cluster, were required to isolate. When individuals who tested positive obtained PCR tests which were negative, concern hit the headlines that the proportion of LFT-positives who were FPs may be high, as the RSS Covid-19 Taskforce had forewarned (Bird, 2021; RSS Covid-19 Taskforce, 2021).

The proportion of LFT-positives that are RT-PCR negative depends on the prevalence. Using the figures from Liverpool (Example 8) for the performance of the Innova test shows that the PPV ranges from 80% if 1 in 100 is infected, to 29% if 1 in 1,000 is infected and 4% if only 1 in 10,000 is infected (Deeks, 2021). The exact prevalence of asymptomatic infection in secondary school children was unknown but estimates from ONS Community Infection Survey suggested an overall prevalence of 0.4%, which would include symptomatic cases and those post-infection with residual inactive virus so that the RSS Covid-19 Taskforce anticipated that half would be asymptomatic infections.

	Covid	No Covid	Total	Percentage with infection	
Scenario A—1 in 100 pupils have Covid-19 infection					
LFT+	4,000	990	4,990	LFT+	80.2%
LFT−	6,000	989,010	995,010	LFT−	0.6%
Total	10,000	990,000	1,000,000	Overall	1.0%
Prevalence	1.00%				
Sensitivity of LFT	40.00%				
Specificity of LFT	99.90%				
Scenario B—1 in 1,000 pupils have Covid-19 infection					
LFT+	400	999	4,990	LFT+	28.6%
LFT−	600	989,010	995,010	LFT−	0.06%
Total	1,000	999,000	1,000,000	Overall	0.10%
Prevalence	0.10%				
Scenario C—1 in 10,000 pupils have Covid-19 infection					
LFT+	40	1,000	1,040	LFT+	3.8%
LFT−	60	998,900	998,960	LFT−	0.006%
Total	100	999,900	1,000,000	Overall	0.01%
Prevalence	0.01%				

Data for the first 2 weeks of testing were made available at the end of April 2021 and showed that 1,050 of the 2,304 positive lateral flow test results had been verified by RT-PCR (in contravention of Government

recommendations) (Department of Health & Social Care, 2021). As predicted, the proportion of positive LFTs that were false was high, with 605 negative and 428 positive (17 tests were void): a PPV of only 41%. Regardless of the RT-PCR result, students, their families and their class-contacts had to isolate for 10 days. Routine confirmatory PCR testing for all LFT-positives was reintroduced in England 3 weeks after the mass testing in schools commenced.

6.8 Advanced study design issues

6.8.1 Comparisons of tests

Identification of the better performing tests is best obtained from direct head-to-head comparisons of tests undertaken in the same individuals or between randomized groups (Takwoingi et al., 2013) (Section 5.2.5).

For antibody tests, beyond the work involved in using multiple tests on each sample, these studies are logistically straightforward as it is relatively easy and uncontroversial to obtain from participants a large enough venous blood sample to run multiple tests (Public Health England, 2020). Head-to-head comparison of point-of-care antibody tests in the appropriate setting requires participants to produce enough finger prick blood for multiple lateral flow devices, which is a greater patient burden and limits the number of devices which can be tested simultaneously. This led to many head-to-head comparisons of point-of-care antibody tests being done with samples taken from the same patients at multiple time points (Flower et al., 2020; Moshe et al., 2021); or performed in laboratory settings on venous blood rather than in the use case setting.

Given the dependence of rapid antigen tests on viral load, there are high risks that between study comparisons of test performance could be confounded if there are differences in viral levels between samples in different studies. Robust comparative studies of antigen and molecular tests were needed: there were far fewer than for antibody tests. The Cochrane review of rapid tests identified only three out of 48 studies directly comparing antigen tests: most were done in laboratory rather than clinical settings (Dinnes et al., 2021). Many were done using alternative samples to swabs, such as viral transport media (e.g. Pickering et al., 2021) which may be outside the approved specimen types.

Where it is not possible to compare all tests in all participants, experimental designs can be considered. Suppose that patients can supply enough finger-prick blood for comparison of three out of four index tests (A, B, C, and D). Then each patient-volunteer can be randomized to which trio of comparisons their donation will be used for: ABC or ABD or ACD or BCD together with the reference standard. Sufficient blood for comparison of a pair out of four tests would entail randomization to one of six possible pair-wise comparisons: AB or AC or AD or BC or BD or CD. Good practice would also involve randomizing swab-order as it may matter.

6.8.2 Using efficient designs

The prevalence of SARS-CoV-2 infection was generally less than 0.5% but with a fourfold increase or decrease also observed during pandemic waves. Hence, prospective studies needed to screen significant numbers of individuals to be able to identify adequate numbers with infection to be able to estimate sensitivity with adequate precision. As the most expensive component of a test evaluation study is ascertaining disease status using PCR tests, designs which reduce the numbers of PCR tests done in those most likely not to have the infection will be more efficient (Holtman et al., 2019). A straightforward way of doing this is to test all who are positive on the rapid antigen test and a random sample of those who test negative (see Example 15). Whilst predictive values can be estimated directly, as they are estimated within groups sampled with the same probability, estimation of sensitivity, specificity and prevalence requires weighting according to the inverse of the sampling probability.

Example 15 (Use of sampling for efficient designs). One study used sampling to provide efficient estimates of the performance of the Innova lateral flow assays in the UK (see Example 8). The University of Birmingham study of testing in students from December 2020 verified 720 Innova tests (90 tests per day for

8 days) with RT-PCR: all Innova test positives (2/2) and 718 test negatives which was a 10% sample from the total 7,187 tested. Estimation of sensitivity, specificity, and prevalence was undertaken by weighting according to the inverse of the sampling probabilities yielding estimates of 3% (95% CI: 1 to 16), 100% (95% CI: 99.5 to 100), and 0.9% (95% CI: 0.4 to 1.9), respectively (Ferguson et al., 2021).

The same analytical issue arises when individuals are allowed to choose whether or not to get a confirmatory RT-PCR test. For example, a study of an (unnamed) lateral flow antigen test in Wales observed a difference in the rate of confirmatory PCR testing of 48% in those with LFT positives compared to 2% in those with LFT negatives (Cwm Taf Morgannwg Test Trace Protect Service, 2021). As those self-selecting for confirmatory PCR testing, particularly amongst those who were testing negative, are unlikely to be a representative sample, valid estimates of sensitivity and specificity cannot be obtained from the data collected, and no analysis of test performance is included in the report. Sampling which ensures the selected groups are representative, such as random sampling, is required.

6.9 SARS-CoV-2 surveillance studies

DHSC-funded surveillance studies of SARS-CoV-2, even when designed in a manner that obviates knowledge by the diagnostic laboratory about the personal identifying information of those who participated in surveillance, were obliged to disclose to NHS Test and Trace personal identifying information about all participants who provide a swab for PCR-testing. In addition, for those whose PCR test is positive or indeterminate, their phone number and email address were reported to NHS Test and Trace.

Hence, unlike in the HIV and HCV epidemics, DHSC-funded surveillance during the Covid-19 pandemic, which involved testing for SARS-CoV-2 antigen, was not allowed to be anonymized as disclosure of personal and private information was mandated. Anonymity delivered high volunteer rates and frankness in HIV and HCV testing.

6.10 Assessing test strategies and their impact

In the Covid-19 pandemic, there were uncertainties about adherence to self-isolation and how well LFTs performed—singly, in combination, and when used serially. For highly infectious diseases, individuals' test-related behavioural choices impact others—at home, at work, or in wider society—for whom different outcomes (whether infection or isolation) have individualized trade-offs. Tests that give FP results can lead to unnecessary self-isolation, and risk harming quality of life and the economy. Tests that give FN results can lead to false impressions of safety and inappropriately put others at risk of infection. Modelling test strategies helps to gauge potential outcomes but cannot account fully for the complexity of interactions between testing, test errors, intervention, and human behaviours.

As with drugs and vaccines, the best evidence on the impact of testing strategies for patients and public health can be obtained from randomized controlled trials (RCTs). As such, RCTs compare strategies that combine the use of a test with intended subsequent management according to test result(s) and measure outcomes according to randomized assignment (Ferrante di Ruffano, Hyde et al., 2012). Several governments promoted the use of RCTs to assess whether testing strategies could be useful in managing the Covid-19 pandemic.

Five examples of studies of test strategies are summarized in Table 1. Two city-based RCTs (Barcelona (Revollo et al., 2021) and Paris (Delaugerre et al., 2022)) and an observational study (Liverpool (Burnside et al., 2024)) assessed freedom to attend mass cultural events during the Covid-19 pandemic; two ambitious large RCTs evaluated the use of serial LFTs to enable release from self-isolation by the primary contacts of an index case. The schools-based cluster RCT recruited 201 schools with over 200,000 students and 25,000 staff. The second, a pseudo-RCT, consented to over 50,000 primary contacts of index cases notified from NHS Test and Trace (Department of Health & Social Care, 2020c); identified those that became secondary cases; and monitored whether their contacts then became tertiary cases.

Table 1. Summary of experiments on test strategies to evaluate more permissive public policies

Studies	Barcelona (Revollo et al., 2021)	Liverpool (Burnside et al., 2024)	Paris (Delaugerre et al., 2021)	Schools (Young et al., 2021)	NHSTT[a] Contacts (Love et al., 2022)
Intervention dates	Dec 12 2020	Apr 28–May 2 2021	May 29 2021	Mar18–May 4 2021	Apr 29–Jul 28 2021
Setting	Live indoor concert	2 nightclubs, outdoor music festival, business conference	Live indoor concert	Students (≥11y) and staff secondary schools and colleges	Adult contacts identified by NHSTT of Covid-19 cases
LFT	1×Panbio by healthcare staff 24 hr before	1×Innova by supervised testing 36 hr before	1×Standard Q Covid-19 Ag Test healthcare staff 72 hr before	5×Orient Gene daily self-swab tested by staff	7×Innova daily self-swabbed self-test
Number ineligible because LFT positive	0 of 1,140	1 of 12,351	1 of 6,968	N/A	N/A
Underlying population weekly incidence	0.13% on day of event	0.0136% on day of event	0.25% Sep 2020	N/A	N/A
Study design	LFT-ve participants randomized	LFT-ve participants (no comparison)	LFT-ve participants 2:1 randomized	201 schools cluster randomized	Contacts randomized. Both groups receive PCR-self-swab to be used on Day 1; and returned by post.
Intervention strategy (n participants, k cluster)	$n = 523$ Randomized to attend 1-day event	$n = 12{,}256$ Attend 1-day event	$n = 4{,}451$ Randomized to attend 1-day event	$n = 5{,}797$ k = 102 Randomized daily LFT + ve (7 days)	$n = 26{,}123$ Randomized daily LFT + ve, PCR on days 1 and 7 (7 days)
Control strategy (n participants, k cluster)	$n = 524$ Randomized to not attend the event	(no comparison)	$n = 2{,}227$ Randomized to not attend the event	$n = 4{,}400$ $k = 99$ Randomized isolated (10 days)	$n = 23{,}500$ Randomized isolated, PCR on day 1 (10 days)
Outcome PCR tests day 0/1	PCR day 0 screening swab by healthcare (100% [b]complete)	PCR day 1 submit self-sample by post (18% [b]complete)	PCR day 1 self-sample saliva at event (100% [b]complete)	(No data presented on Day 0/1)	(No data presented on Day 0/1)
Outcome PCR tests after strategy	PCR day 8 visited swab by healthcare worker (97% [b]complete)	PCR day 5 to 7, submit self-sample by post (37% [b]complete)	PCR day 7 self-sample saliva sent by post (88% [b]complete)	PCR self-sample swab sent to NHSTT if LFT positive or symptomatic	PCR self-sample swab sent to NHSTT if LFT positive or symptomatic

(*continued*)

Table 1. Continued

Studies	Barcelona (Revollo et al., 2021)	Liverpool (Burnside et al., 2024)	Paris (Delaugerre et al., 2021)	Schools (Young et al., 2021)	NHSTT[a] Contacts (Love et al., 2022)
PCR positive intervention strategy cases	0 of 465	8 of 2,214	8 of 3,917	740 cases (rates based on 1,197,107 person-weeks)	327 tertiary cases of 5,191[c] secondary contacts (6.3%)
PCR positive control strategy cases	2 of 495	(No comparison)	3 of 1,947	657 cases (rates based on 1,111,791 person-weeks)	391 tertiary cases of 5,219[c] secondary contacts (7.5%)
Effect size comparing cases	Risk difference −0.15% (95% CI: −0.72% to 0.44%)		Risk difference 0.05% (95% CI: −0.26% to 0.28%)	Rate ratio 1.05 (95% CI: 0.71 to 1.55)	Risk difference −1.2% (95% CI: −2.3% to −0.2%)
Effect size for absences				Rate ratio 0.83 (95% CI: 0.54 to 1.26)	

[a]NHSTT = NHS Test and Trace.
[b]Complete indicates the number of participants who had complete data for the PCR test.
[c]NHSTT study selected for analysis tertiary cases who were close contacts of the RCT's secondary cases those who had tested PCR-positive after having been an identified and randomized close contact of a primary case. The NHSTT RCT consented and randomized close contacts of primary cases. The close contacts of secondary cases were neither consented nor randomized and so had to heed the self-exclusion rules that applied at the time.

6.10.1 Freedom to attend mass cultural events

Three studies reported the use of a negative LFT to identify individuals who can safely take part in mass cultural events: two RCTs undertaken for indoor concerts, and an observational study of attendance at one of four venues (two nightclubs, an outdoor music festival and a business conference).In the Barcelona and Paris RCTs, and the observational study in Liverpool, healthcare staff assessed all participants using an LFT on the same day as the concert (Barcelona) or 2 (Liverpool) or 3 days (Paris) before. Only one participant in Paris, none in Barcelona, and one in Liverpool were found to be LFT-positive at screening and hence not allowed to attend the event. Those who were LFT-negative were randomized—in the ratio 2:1 in Paris and 1:1 in Barcelona—to either attend the event (event group) or follow standard lockdown rules (comparison group). Participants were also assessed with a PCR test on the day of the event; and post-event (on Day 7 or 8). The Paris study provided participants with prepaid envelopes to return saliva swabs, whereas the Barcelona study had study staff visit participants for throat swabbing. Follow-up was 99% (event group) and 94% (control group) in Barcelona, 87% (event group) and 88% (control group) in Paris, and 37% in Liverpool (event group).

Infection rates were very low in both studies. In Paris, only 11 positive PCR tests were detected (8/3917, 0.20% positive in the event group vs. 3/1947, 0.15% positive in the control group). See Table 1 for further details for Barcelona and Liverpool. The Liverpool study was nonrandomized. Poor compliance with post-event testing magnified uncertainty about the risk of infection and safety of events. Safety considerations might change if the pre-event prevalence of infection were high.

6.10.2 Contacts' release from self-isolation

Table 1 summarizes the design and findings of the two RCTs in England that evaluated whether contacts of people who had tested PCR-positive could safely avoid isolation by using repeated LFTs. In the first, 201 secondary schools or colleges were the unit of randomization in a cluster-RCT (Young et al., 2021). The second is essentially randomized by household of those who gave consent among the close contacts whom an index case had notified to NHS Test and Trace (Love et al., 2022). Randomization in both studies occurred prior to both testing and isolation of participants.

The Daily Contact Testing (DCT) schools' cluster-RCT evaluated school absence and the transmission of SARS-CoV-2 in the randomized schools. Students and staff who were PCR positive were isolated at home for 10 days; their close contacts were identified according to the national policy. Cluster randomization allocated schools to be either intervention or control schools: 80% of 201 randomized schools participated actively. Schools allocated to the DCT strategy allowed contacts to attend school provided they received daily negative LFTs (for 7 days); in the control schools, contacts were directed to isolate at home for 10 days. Schools reported absences of all students and staff for Covid-19-related reasons. Symptomatic PCR-positive cases in each school were identified from NHS Test and Trace.

In the DCT schools, Covid-19-related absence was slightly lower than in the control group of self-isolation schools: a rate ratio of 0.83 (95% CI: 0.54 to 1.26). Infection rates were generally low and similar in the self-isolation schools compared to the control group: a rate ratio of 1.05 (95%CI: 0.71 to 1.55). Confidence intervals for both outcomes were wide such that the impact of the policy was unclear. The sensitivity of the chosen LFT (Orient Gene) was 53% (32/60; 95% CI: 40% to 66%) as a single screening test.

The NHS Test and Trace DCT trial pseudo-randomized consented primary contacts ($n = 54{,}923$) of index cases identified by NHS Test and Trace with a view to measuring the number of primary contacts who became secondary cases ($n = 4{,}694$) and the number of their contacts (i.e. nonrandomized secondary contacts, $n = 10{,}410$) who became tertiary cases ($n = 718$). Only primary contacts were consented and randomized: so that the usual self-isolation rules pertained to secondary contacts.

Participants in both groups were asked to take a self-sample for PCR on the day of kit arrival and return it by post. The DCT primary contacts were provided with seven Innova LFTs and asked to perform their first LFT on the day of kit arrival (Day 1) and daily thereafter; and to complete a second self-swab PCR test on the last day (Day 7) or earlier if an LFT was positive. The DCT

primary contacts were only required to isolate if they had a positive LFT or symptoms. The self-isolation primary contacts completed a single PCR test on Day 1 and were isolated for 10 days. All participants were legally required to self-isolate for 10 days if PCR-positive.

The initial randomization was adapted so that all consented close contacts within a given household were assigned to the same intervention as the earliest consented and randomized household-member: the reassigned counts of 28,757 (52.4%) to DCT vs. 26,166 (47.6%) to self-isolation were very significantly deviant from 50:50. The random assignments of the first-consented primary contacts are not clear; of the 10,410 secondary contacts, 74% had been identified by secondary cases who were randomized as the first participant in a household.

The DCT kit, which provided for two PCR swabs, led to more PCR tests being reported in the DCT group (1.53 PCR tests per person) than in the isolation group (1.13 PCR tests per person). Intrinsic to 'intention to manage,' differential provision of PCR swabs may have enhanced PCR detections in DCT-assigned primary contacts. The report does not tabulate or investigate PCR results by random assignment and PCR timing.

The analysis focused on the un-consented and nonrandomized 10,410 secondary contacts (only 1,214 of them external secondary contacts), whom the 4,694 secondary cases had identified to NHS Test and Trace. As to transmission risk, 718/10,410 secondary contacts became PCR-positive (7%), i.e. 718 tertiary cases. The percentage of tertiary cases derived from the DCT group's secondary contacts was lower than from the isolation group: absolute risk difference of −1.3% (95% CI −2.3 to −0.2), showing noninferiority at the pre-stated margin of 1.9%. The magnitude and precision of the difference reduced when the analysis was restricted to the first-consented primary contact in each household: −0.8% (95% CI −2.0 to 0.4). The rationale for the pre-stated margin of 1.9% was not given clearly.

Pragmatic RCTs have an important scientific role in assessing the impact of testing and management strategies on patient outcomes. Randomized controlled trials of test-intervention strategies need to be carefully planned to minimize the risk of bias in their design and execution, and to compute appropriately the sample sizes required to detect the important differences in outcomes that testing strategies can create.

6.10.3 Methodological and logistical challenges

Randomized controlled test-intervention strategy trials are recommended (Horvath et al., 2014) *to compare test strategies in alignment with subsequent management* and measure outcomes to assess whether a test strategy leads to more good than harm. However, test-intervention RCTs are rarely used (Ferrante di Ruffano, Davenport et al., 2012) outside of those evaluating screening strategies (Department of Health, 2015; Welch & Black, 1997), and face methodological and logistical challenges (Bossuyt et al., 2000; Ferrante di Ruffano, Dinnes et al., 2017) particularly with diseases which naturally have low prevalence (Holtman et al., 2019).

Randomized controlled trials can provide the strongest evidence of impact, provided that they are not at risk of bias, whether selection, performance, detection, or attrition (Higgins et al., 2023). Test-intervention RCTs face particular methodological and logistical challenges: (a) requiring exceptionally large sample sizes as the impact can only occur in the small subset of participants who receive different interventions based on test results; (b) identifying and measuring the multiple outcomes that affect participants from test to intervention; (c) accounting for individuals making their own choices as to whether to be tested and to follow intervention pathways (Ferrante di Ruffano, Hyde et al., 2012); and, for infectious diseases, (d) clustering of outcomes (in this case, transmission) due to nonindependence of participants who share the same household/or workplace and infect each other therein (Weijer et al., 2021).

The UK's two cluster RCTs during Covid-19 demonstrate methodological innovation to balance release from quarantine against the rate of infectious disease transmissions. Future research must be considered a priority to investigate the occurrence of transmission events that may be unidentified (because infector and infectee are not known to each other) within households, schools, or workplaces.

For all five studies in Table 1, the sample sizes required to obtain an adequately precise estimate of the difference in strategies were challenging, both in terms of identifying the number of participants required and completing their follow-up.

The sample size requirements for a randomized controlled test-intervention strategy are challenging for all diseases. Future research should investigate the use of alternative study designs, including the early use of comparative randomized accuracy studies (Takwoingi et al., 2013), particularly for more complex strategies such as repeated testing.

There is a place for innovative adaptive study-designs which combining accuracy with effectiveness objectives within the same RCT cohort. Similarly, there are opportunities to investigate competing durations of quarantine of contacts in the household of an index case; and, separately, for external close contacts—at the same time as commencing studies of effectiveness and impact.

7. Regulation

Diagnostic tests are considered devices from a regulatory standpoint. In the UK, MHRA is responsible for regulating medicines, medical devices, and blood components (Medicines & Healthcare products Regulatory Agency, 2018). However, there are significant differences in the oversight of medicines and devices. Medicines regulation is financed through manufacturers' licence fees, while the DHSC finances the regulation of devices, including IVDs.

Before 2022, the typical method of certification of IVDs involved placing a Conformité Européene (CE) mark on the product, with the manufacturer being primarily accountable for this, rather than the regulator. The 1998 Directive (Box 5) outlines the responsibilities of the manufacturer and the bodies they inform (European Parliament & Council, 1998).

Box 5: The 1998 Directive statement about responsibilities.

'(22) whereas, since the large majority of such devices do not constitute a direct risk to patients and are used by competently trained professionals, and the results obtained can often be confirmed by other means, the conformity assessment procedures can be carried out, as a general rule, under the sole responsibility of the manufacturer; whereas, taking account of existing national regulations and of notifications received following the procedure laid down in Directive 98/34/EC, the intervention of notified bodies is needed only for defined devices, the correct performance of which is essential to medical practice and the failure of which can cause a serious risk to health.'

The new European Union Regulation (EU) 2017/746 on IVD medical devices was implemented in EU countries on 23rd May 2022 (European Parliament & Council, 2017). This Regulation includes new compulsory arrangements for transparency; and limits self-certification. Current IVD device regulations vary by risk class. Low-risk devices can self-certify, but IVDs for infectious diseases require public documents summarizing the safety and performance of the device as well as (regulatory) Notified Body involvement.

The EU Regulation emphasizes the risk of misdiagnosis in the use of IVDs and assigns responsibilities to the Notified Body for independent verification, audit of quality management, and assessment of technical documentation. The sale of IVD device in the EU market requires compliance with European laws. The MHRA is defining new regulations for Great Britain (Medicines & Healthcare products Regulatory Agency, 2023), which are expected to harmonize domestic laws with EU regulations. However, important areas still require improvement, such as granting public access to the manufacturer's data underlying IVD registration.

One approach adopted by MHRA and other regulators during the Covid-19 pandemic was the use of Target Product Profiles (TPPs) (Medicines & Healthcare products Regulatory Agency, 2020b). A TPP outlines the characteristics of a product intended to treat specific diseases with specifications on intended use, target populations, and safety and performance-related attributes that manufacturers and regulators need to address. While the current TPPs outline the necessary properties for a reference standard to establish test performance, they do not specify whether the reference standards meet these criteria. Achieving greater standardization of research could be possible by achieving consensus on suitable reference standards and incorporating them into future TPPs.

Currently, tests which are to be administered by members of the public rather than healthcare professionals do require evaluation by a Notified Body to obtain a CE marking that allows self-use. During the Covid-19 pandemic, regulators provided time-limited EUAs for IVDs—a route to approval separate from the CE mark. In the UK, the exceptional use authorized for some IVD devices on a time-limited basis was for use by the public for self-testing, bypassing the more strenuous approval process usually required. Devices granted EUA were sold to the NHS and for use in social care.

Devices authorized by this route were publicly listed (Medicines & Healthcare products Regulatory Agency, 2020c), whereas those approved by usual routes have not been. Outside of EUAs, a confidentiality clause ruled out a more comprehensive public register of devices, and it has been cited as a reason for not responding to Freedom of Information Requests in the UK. The confidentiality clause was revised by the new EU regulatory framework. In April 2017, the EU brought into force a new Regulation on IVD Medical Devices, which stated: 'a fundamental revision [to the law] is needed to establish a robust, transparent, predictable and sustainable regulatory framework for IVD medical devices which ensures a high level of safety and health whilst supporting innovation.'

8. Information to be in the public domain

8.1 Ensuring scientific integrity

Accurate, comprehensive information about the evaluation and performance of diagnostic tests is essential: to allow the public and clinicians to make informed decisions on being tested and on interpreting test results appropriately; to enable policymakers to decide on testing strategies and the procurement and deployment of tests; and for researchers to be fully informed about existing research and to plan appropriately the next studies.

The importance of well-organized, transparent reporting of all stages of research has been established for RCTs of interventions, with many aspects now enforced by the leading medical journals (De Angelis et al., 2005). Transparent reporting of diagnostic test evaluations is equally critical to ensure scientific integrity as for RCTs. Aspects to consider encompasses:

- Prospective public registration to ensure all research studies can be identified: to prevent publication bias (Simes, 1986).
- Prospective publication of study protocols and statistical analysis plans to document the original study-design and pre-specified outcome measures: to distinguish pre-specified analyses from potential data-driven analyses (Chan & Hróbjartsson, 2018).
- Peer review of protocols and study reports: to validate the science and enhance the quality of reports (Yordanov et al., 2018).
- Timely open-access publication (Dwan et al., 2013).
- Access to data sets: to enable study findings to be confirmed and data presentation to be harmonized (Taichman et al., 2017).

Test evaluation studies face similar concerns about selective publication and data-driven analyses as clinical trials did. Therefore, the same principles should be endorsed. Setting standards for prospective public registration of diagnostic test evaluations will require international efforts, as previously for trials (Gülmezoglu et al., 2005). However, there already exist proposals for pre-registration of all prospective studies (Naudet et al., 2024), including test evaluations.

8.2 What the public, patients, and clinicians need to know?

Those providing testing to the public and patients have a responsibility to ensure that individuals offered testing can make an informed choice, and that they appreciate potential downsides as well as benefits of testing. In the past, the promotion of screening may have played down potential harms, as clinicians were concerned that 'if you tell people the whole truth, getting them into the screening programme will somehow be jeopardized' (Science & Technology Committee Inquiry into National Health Screening, 2014). The same perspective can affect testing for infectious diseases, particularly where a key benefit may be as much to the community as simply to the

individual being tested. With few exceptions, mandatory testing is contrary to the UK public interest.

Informed choice entails providing clear, unbiased information to enable participants to assess the offer of testing and decide whether to accept or decline it. Choices are influenced by personal circumstances and values, and individuals differ in how they balance benefits and risks.

The public and patients need trustworthy information about the chances that they might obtain FP and FN results, the consequences which these could create, and be able to think through the actions and decisions that they would take and make given positive or negative results. Presentation of accessible information and data relevant to the intended use of the test, tailored to account for differences in disease prevalence, with results presented in formats that the public understands is thus essential (see Examples 13 and 14) (Spiegelhalter & Masters, 2021).

Particularly important is the understanding that a second, often different, test may be required to determine the presence of disease after a screening test has signalled a need for further checking. Equally important is to warn that a screening test may fail to alert. Hence, explaining data in terms of probabilities conditional on test results (e.g. predictive values) ensures both that disease prevalence is accounted for and that individuals properly comprehend the implications of positive and negative test results.

For example, for lateral flow antigen tests initially used during Covid-19 in the UK, there were particular issues in the explanation of the poor predictive value of negative test results due to the low sensitivity of the tests. As explained in Section 6.7, negative test results did not rule out infection or infectiousness. It was essential that the public was aware of this, as misinterpretation of a negative test result as indicating an individual is safe and does not have the infection could lead to disinhibition, greater risk-taking, and hence increased transmission.

8.3 What policymakers need to know

During the pandemic, decisions had to be hedged about which tests to purchase because the available tests had to be evaluated in more limited contexts of use than when they were eventually applied. As with vaccines, it may be prudent for governments to diversify their portfolio of test purchases and make changes in the light of new evidence.

The approach which was adopted and endorsed by the UK Government for review of screening programmes contains processes and information of relevance to the more general use of tests, particularly for consideration in the commissioning of mass testing which can occur with infectious diseases. The Government Response to the Science and Technology Select Committee review of National Health Screening (Department of Health, 2015) stated that 'screening programmes are only introduced when there is sufficient evidence that the benefits outweigh any potential harms, and that people are given all the facts before making an informed decision to take up an offer of screening'. Potential harms include 'giving a negative result when the results should be positive (a FN result) thereby missing the correct diagnosis; or giving a positive result when the result should have been negative (a FP result). This may result in stress for the individual and possible follow-up treatment that is unnecessary'. For most screening, it is sufficient to consider the individual perspective. For infectious diseases, there is a public health dimension as well.

Policy decisions about the introduction of tests, particularly for mass testing, need to consider evidence of the likely benefits and potential harms and assess whether the balance is favourable and provides appropriate value for money. The absolute numbers of FP and FN will change with the prevalence of disease, altering the profile of benefits and harms. At higher disease prevalence, more cases will be detected although some positive tests will be FP. As infection levels drop, fewer cases will be detected, whereas the likely harms through FP will remain the same. Hence, the benefit-to-harm ratio will become less favourable, as will the costs and resources required to detect each case.

Real-time monitoring of test performance and disease incidence is therefore essential to ensure that testing stops before the harms outweigh the benefits, and that processes (such as confirmatory testing) are in place to mitigate the risks from wrong initial test results (see Example 14). Other aspects of testing, such as the failure rate of tests, the time taken to obtain a test and receive results, the acceptability of the sampling approach, and the ease of access to testing, will all impact policy decisions.

Tests alone do not improve patient and population outcomes; it is the interventions that follow that create benefit. Therefore, testing implementation needs to be linked to the implementation of the interventions required for patient and population benefits to be realised and to ensure that the information provided to those being tested leads them to undertake the correct actions. For example, ensuring individuals positive for SARS-CoV-2 infection can be isolated was essential for preventing onward transmission of infection.

Policymakers thus require access to all study findings, which can usefully be summarized in systematic reviews that identify, appraise, and synthesize all relevant evidence and assess the strength of the evidence (Deeks et al., 2023). Creation of this evidence resource requires all studies to be published with full information given about their methods and findings. During a pandemic, it is essential that systematic reviews are undertaken in a timely manner and updated as new information becomes available. The availability of preprints from online archiving services revolutionised the ability to achieve this.

Separate evidence reviews are required for each intended use, each assessing the strength of the evidence, noting the findings and the inherent uncertainty in the estimates (using confidence intervals or equivalent), consistency of findings, applicability of the evidence to the intended use, and confidence that the findings are based on complete data and not at risk of bias.

To be of maximal use, study reports, both primary test evaluations and systematic reviews, need to be fully reported to enable critical appraisal of their methodology and findings. Reporting guidelines for primary studies (STARD) (Bossuyt et al., 2015) and systematic reviews (PRISMA-DTA) (McInnes et al., 2018; Salameh et al., 2020) should be followed to ensure that all relevant details are included.

8.4 What researchers and study participants need to know

Scientific integrity can be compromised when research studies are not made public or key details are omitted, either for commercial reasons or academic interests. For example, due to the commercial confidentiality clauses in the contracts with the manufacturers, DHSC-sponsored evaluations of point-of-care antibody tests did not name the nine test kits evaluated (Adams et al., 2020). This leads to research waste (Glasziou et al., 2014) as others may study the same tests unaware of the already completed work. Pre-registration of studies, publication of protocols, and timely publication of results are essential to ensure that the research efforts are directed appropriately. Preparation for a pandemic should involve designing and producing protocols for generic studies ahead of time, as has happened for influenza (Goodacre et al., 2015).

Many test evaluations are undertaken as service evaluations rather than as research, which risks that aspects of a research study, such as open protocol and Good Clinical Practice Guidelines that protect patients, could be bypassed. This is particularly inappropriate where studies require extra information or procedures (such as reference standard tests) to be undertaken or benefit from follow-up of patients to maximize the accuracy of the reference standard or assess sequelae.

Acknowledgements

Members of the RSS Covid-19 Taskforce and Stian Westlake peer-reviewed the report. Secretariat support has been provided by Olivia Varley-Winter from the Royal Statistical Society. Hayley Walton, Bethany Hillier, and Simon Baldwin at the University of Birmingham undertook detailed checking of the final document. We are grateful to the staff at the MHRA for providing information on current regulatory issues.

Conflicts of interest: Professor D.A. is a co-investigator on the REACT studies. Professor J.D. was a member of the MHRA IVD external advisory group; consultant to the WHO Essential Diagnostics List; and led the Cochrane Covid-19 diagnostic test reviews team; co-investigator of the Birmingham University evaluation of the Innova test. He leads the NIHR Birmingham BRC Data, Diagnostics and Decision-making theme; chief methodological editor for Cochrane's diagnostic test accuracy systematic reviews; Chief Statistics Editor at the BMJ; co-applicant on MRC, NIHR, and CRUK grants on test evaluation. He is a member of the RSS Covid-19 Taskforce and was a member of the RSS/DHSC Panel on NHS Test and Trace. Professor S.B. is a member of RSS Covid-19 Taskforce; chair of its RSS/DHSC Panel on NHS

Test and Trace; member of NHS Test and Trace/Public Health England Testing Initiatives Evaluation Board; grant-funded contribution to the design and analysis of unlinked anonymous surveillance study in London of SARS-CoV-2 antibodies at antenatal booking visits throughout 2020. Holder of GSK shares. Professor S.E. has no conflicts of interest. Professor R.P. is lead for the Methods Theme of the NIHR Oxford Medtech and In-Vitro Diagnostics Co-operative as well as the NIHR Oxford and Thames Valley Applied Research Collaborative (ARC). Professor Y.T. is a co-convenor of the Cochrane Screening and Diagnostic Tests Methods Group; member of the Cochrane Editorial Board; and co-investigator on MRC and NIHR funded test evaluation projects.

Funding

J.J.D. and Y.T. are supported by the NIHR Birmingham Biomedical Research Centre. D.A. is supported by the NIHR Imperial Biomedical Research Centre. R.P. was supported by the NIHR Community Healthcare MedTech and In Vitro Diagnostics Co-operative and the NIHR Oxford and Thames Valley Applied Research Collaborative (ARC).

Data availability

The material in this report is based on documents cited in the references. No additional data are available.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

References

Abbott Diagnostics. (2020). *ID Now Instructions for Use. IN 190000 Rev. 7 2020/09*. Date accessed May 21, 2021 https://www.fda.gov/media/136525/download

Abingdon Health. (2021a). *UK Covid-19 rapid antibody tests approved for professional use*. Date accessed May 19, 2021. https://www.abingdonhealth.com/uk-Covid-19-rapid-antibody-tests-approved-for-professional-use/

Abingdon Health. (2021b). *ABC-19 response to media*. Date accessed May 21, 2021. https://www.abingdonhealth.com/news/abc-19-response-to-media/

Adams, E. R., Ainsworth, M., Anand, R., Andersson, M. I., Auckland, K., Baillie, J. K., Barnes, E., Beer, S., Bell, J. I., Berry, T., Bibi, S., Carroll, M., Chinnakannan, S. K., Clutterbuck, E., Cornall, R. J., Crook, D. W., de Silva, T., Dejnirattisai, W., Dingle, K. E., … Whitehouse, J. (2020). Antibody testing for COVID-19: A report from the National COVID Scientific Advisory Panel. *Wellcome Open Research*, *5*, 139. https://doi.org/10.12688/wellcomeopenres.15927.1

Allan, C., Joyce, T. J., & Pollock, A. M. (2018). Europe's new device regulations fail to protect the public. *BMJ (Clinical Research Ed)*, *363*, k4205. https://doi.org/10.1136/bmj.k4205

Alonzo, T. A., Pepe, M. S., & Moskowitz, C. S. (2002). Sample size calculations for comparative studies of medical tests for detecting presence of disease. *Statistics in Medicine*, *21*(6), 835–852. https://doi.org/10.1002/sim.1058

Arevalo-Rodriguez, I., Buitrago-Garcia, D., Simancas-Racines, D., Zambrano-Achig, P., Del Campo, R., Ciapponi, A., Sued, O., Martinez-Garcıa, L., Rutjes, A. W., Low, N., Bossuyt, P. M., Perez-Molina, J. A., & Zamora, J. (2020). False-negative results of initial RT-PCR assays for COVID-19: A systematic review. *PLoS One*, *15*(12), e0242958. https://doi.org/10.1371/journal.pone.0242958

Armbruster, D. A., & Pry, T. (2008). Limit of blank, limit of detection and limit of quantitation. *The Clinical Biochemist Reviews*, *29*(Suppl 1), S49–S52.

Bachmann, L. M., Puhan, M. A., ter Riet, G., & Bossuyt, P. M. (2006). Sample sizes of studies on diagnostic accuracy: Literature survey. *BMJ (Clinical Research Ed)*, *332*(7550), 1127–1129. https://doi.org/10.1136/bmj.38793.637789.2F

Banoo, S., Bell, D., Bossuyt, P., Herring, A., Mabey, D., Poole, F., Smith, P. G., Sriram, N., Wongsrichanalai, C., Linke, R., O'Brien, R., Perkins, M., Cunningham, J., Matsoso, P., Nathanson, C. M., Olliaro, P., Peeling, R. W., & Ramsay, A.; TDR Diagnostics Evaluation Expert Panel (WHO/TDR) 2006 (2006). Evaluation of diagnostic tests for infectious diseases: General principles. *Nature Reviews Microbiology*, *4*(9 Suppl), S21–S31. https://doi.org/10.1038/nrmicro1523

Bigio, J., MacLean, E. L., Das, R., Sulis, G., Kohli, M., Berhane, S., Dinnes, J., Deeks, J. J., Brümmer, L. E., Denkinger, C. M., & Pai, M. (2023). Accuracy of package inserts of SARS-CoV-2 rapid antigen tests: A

secondary analysis of manufacturer versus systematic review data. *Lancet Microbe*, *4*(11), e875–e882. https://doi.org/10.1016/S2666-5247(23)00222-7

Bird, A. G., Gore, S. M., Jolliffe, D. W., & Burns, S. M. (1992). Anonymous HIV surveillance in Saughton Prison, Edinburgh. *Aids (London, England)*, *6*(7), 725–733. https://doi.org/10.1097/00002030-199207000-00017

Bird, S. M. (2021). COVID tests in secondary schools: A statistical cause célèbre. *Significance*, *18*(3), 42–45. https://doi.org/10.1111/1740-9713.01535

Bland, J. M., & Altman, D. G. (1995). Comparing methods of measurement: Why plotting difference ag ONSainst standard method is misleading. *Lancet (London, England)*, *346*(8982), 1085–1087. https://doi.org/10.1016/s0140-6736(95)91748-9

Bochmann, F., Johnson, Z., & Azuara-Blanco, A. (2007). Sample size in studies on diagnostic accuracy in ophthalmology: A literature survey. *The British Journal of Ophthalmology*, *91*(7), 898–900. https://doi.org/10.1136/bjo.2006.113290

Boseley, S. (2020, March 25). UK coronavirus home testing to be made available to millions. *The Guardian*. Date accessed May 17, 2021. https://www.theguardian.com/world/2020/mar/25/uk-coronavirus-mass-home-testing-to-be-made-available-within-days

Bossuyt, P. M., Lijmer, J. G., & Mol, B. W. (2000). Randomised comparisons of medical tests: Sometimes invalid, not always efficient. *Lancet (London, England)*, *356*(9244), 1844–1847. https://doi.org/10.1016/S0140-6736(00)03246-3

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., Lijmer, J. G., Moher, D., Rennie, D., de Vet, H. C. W., Kressel, H. Y., Rifai, N., Golub, R. M., Altman, D. G., Hooft, L., Korevaar, D. A., & Cohen, J. F. (2015). An updated list of essential items for reporting diagnostic accuracy studies. *BMJ (Clinical Research Ed)*, *351*, h5527. https://doi.org/10.1136/bmj.h5527

Buchanan, L., Gamio, L., Leatherby, L., Keefe, J., Koettl, C., & Walker, A. S. (2020). Tracking the White House Coronavirus Outbreak. *The New York Times*. Date accessed October *5*, 2020. https://www.nytimes.com/interactive/2020/10/02/us/politics/trump-contact-tracing-covid.html

Bullard, J., Funk, D., Dust, K., Garnett, L., Tran, K., Bello, A., Strong, J. E., Lee, S. J., Waruk, J., Hedley, A., Alexander, D., Van Caeseele, P., Loeppky, C., & Poliquin, G. (2021). Infectivity of severe acute respiratory syndrome coronavirus 2 in children compared with adults. *CMAJ*, *193*(17), E601–E606. https://doi.org/10.1503/cmaj.210263

Burnside, G., Cheyne, C. P., Leeming, G., Humann, M., Darby, A., Green, M. A., Crozier, A., Maskell, S., O'Halloran, K., Musi, E., Carmi, E., Khan, N., Fisher, D., Corcoran, R., Dunning, J., Edmunds, W. J., Tharmaratnam, K., Hughes, D. M., Malki-Epshtein, L., … Buchan, I. E. (2024). COVID-19 risk mitigation in reopening mass cultural events: Population-based observational study for the UK Events Research Programme in Liverpool City Region. *Journal of the Royal Society of Medicine*, *117*(1):11–23. https://doi.org/10.1177/01410768231182389

Cao, S., Gan, Y., Wang, C., Bachmann, M., Wei, S., Gong, J., Huang, Y., Wang, T., Li, Y., Lu, K., Jiang, H., Gong, Y., Xu, H., Shen, X., Tian, Q., Lv, C., Song, F., Yin, X., & Lu, Z. (2020). Post-lockdown SARS-CoV-2 nucleic acid screening in nearly ten million residents of Wuhan, China. *Nature Communications*, *11*(1), 5917. https://doi.org/10.1038/s41467-020-19802-w

Carey, R. N., Anderson, F. P., & George, H. (2005). *User demonstration of performance for precision and accuracy; approved guidelines–second edition. CLSI document EP15-A2* (Vol. *25*, pp. 1–49). Clinical and Laboratory Standard Institute.

Case, J. B., Bailey, A. L., Kin, A. S., Chen, R. E., & Diamond, M. S. (2020). Growth, detection, quantification and inactivation of SARS-CoV-2. *Virology*, *548*, 39–48. https://doi.org/10.1016/j.virol.2020.05.015

Cassaniti, I., Novazzi, F., Giardina, F., Salinaro, F., Sachs, M., Perlini, S., Bruno, R., Mojoli, F., & Baldanti, F; Members of the San Matteo Pavia Covid-19 Task Force. (2020). Performance of VivaDiag COVID-19 IgM/IgG rapid test is inadequate for diagnosis of COVID-19 in acute patients referring to emergency room department. *Journal of Medical Virology*, *92*(10), 1724–1727. https://doi.org/10.1002/jmv.25800

Centers for Disease Control and Prevention (CDC). (2023). *Lesson 1: Introduction to epidemiology (online) in principles of epidemiology in public health practice third edition an Introduction to applied epidemiology and biostatistics*. U. S. Department of Health and Human Services. https://stacks.cdc.gov/view/cdc/6914/cdc_6914_DS1.pdf

Cevik, M., Marcus, J. L., Buckee, C., & Smith, T. C. (2021). Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) transmission dynamics should inform policy. *Clinical Infectious Diseases*, *73*(Suppl 2), S170–S176. https://doi.org/10.1093/cid/ciaa1442

Chan, A. W., & Hróbjartsson, A. (2018). Promoting public access to clinical trial protocols: Challenges and recommendations. *Trials*, *19*(1), 116. https://doi.org/10.1186/s13063-018-2510-1

Cwm Taf Morgannwg Test Trace Protect (TTP) Service. (2021, March 25). *Evaluation of the lateral flow device testing pilot for Covid-19 in Merthyr Tydfil and the lower Cynon Valley*. Date accessed May 21, 2021. https://cwmtafmorgannwg.wales/Docs/Publications/FINAL_V2_Whole%20Area%20Testing%20Evaluation%20Full%20Report%2020210325.pdf? boxtype=pdf

De Angelis, C. D., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., Kotzin, S., Laine, C., Marusic, A., Overbeke, A. J., Schroeder, T. V., Sox, H. C., & Van Der Weyden, M. B.; International Committee of Medical Journal Editors. (2005). Is this clinical trial fully registered? A statement from the International Committee of Medical Journal Editors. *The New England Journal of Medicine*, *352*(23), 2436–2438. https://doi.org/10.1056/NEJMe058127

Deeks, J. (2021). *Why the school testing regimen needs to change*. UnHerd. Date accessed May 19, 2021. https://unherd.com/thepost/why-the-school-testing-regime-needs-to-change/

Deeks, J., Gill, M., Bird, S., Richardson, S., & Ashby, D. (2021a, January 12). *Covid-19 Innova testing in schools: don't just test, evaluate*. BMJ. Date accessed May 19, 2021. https://blogs.bmj.com/bmj/2021/01/12/covid-19-innova-testing-in-schools-dont-just-test-evaluate/

Deeks, J. J., Bossuyt, P. M., Leeflang, M. M., & Takwoingi, Y. (2023). *Cochrane handbook for systematic reviews of diagnostic test accuracy* (1st ed.). John Wiley & Sons.

Deeks, J. J., Dinnes, J., Takwoingi, Y., Davenport, C., Spijker, R., Taylor-Phillips, S., Adriano, A., Beese, S., Dretzke, J., Ferrante di Ruffano, L., Harris, I. M., Price, M. J., Dittrich, S., Emperador, D., Hooft, L., Leeflang, M. M. G., & Van den Bruel, A. (2020). Antibody tests for identification of current and past infection with SARS-CoV-2. *Cochrane Database of Systematic Reviews*, *6*, Art. No.: CD013652. https://doi.org/10.1002/14651858.CD013652

Deeks, J. J., Raffle, A., & Gill, M. (2021b, January 29). *Re:Covid-19: The accuracy of repeated testing is challenging to model—empirical evaluations seems the only solution. Rapid response to: Covid-19: Controversial rapid test policy divides doctors and scientists*. BMJ. Date accessed May 19, 2021. https://www.bmj.com/content/372/bmj.n81/rapid-responses

Deeks, J. J., Singanayagam, A., Houston, H., Sitch, A. J., Hakki, S., Dunning, J., & Lalvani, A. (2022). SARS-CoV-2 antigen lateral flow tests for detecting infectious people: Linked data analysis. *BMJ (Clinical Research Ed)*, *376*, e066871. https://doi.org/10.1136/bmj-2021-066871

Delaugerre, C., Foissac, F., Abdoul, H., Masson, G., Choupeaux, L., Dufour, E., Gastli, N., Delarue, S. M., Néré, M. L., Minier, M., Gabassi, A., Salmona, M., Seguineau, M., Schmitt, S., Tonglet, S., Olivier, A., Poyart, C., Le Goff, J., Lescure, X., … Tréluyer, J.-M. (2022). Prevention of SARS-CoV-2 transmission during a large, live, indoor gathering (SPRING): A non-inferiority, randomised, controlled trial. *The Lancet Infectious Diseases*, *22*(3), 341–348. https://doi.org/10.1016/S1473-3099(21)00673-3

Department of Health. (2015, January). *Government response to the House of Commons Science and Technology Committee*. Report on National Health Screening. Date accessed May 19, 2021. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/399296/national_health_screening_8999.pdf

Department of Health and Social Care. (2020a, December 1). *Rapid evaluation of OptiGene RT-LAMP assay (direct and RNA formats)*. Date accessed May 19, 2021. https://www.gov.uk/government/publications/rapid-evaluation-of-optigene-rt-lamp-assay-direct-and-rna-formats

Department of Health and Social Care. (2020b, October 6). *Press release. Government invests in UK-developed antibody tests from UK Rapid Test Consortium*. Date accessed May 19, 2021 https://www.gov.uk/government/news/government-invests-in-uk-developed-antibody-tests-from-uk-rapid-test-consortium

Department of Health and Social Care. (2020c). *Guidance. NHS Test and Trace: what to do if you are contacted*. Date accessed May 27, 2020. https://www.gov.uk/guidance/nhs-test-and-trace-how-it-works

Department of Health and Social Care. (2021). *Transparency data. Weekly statistics for NHS Test and Trace (England): 15 April to 21 April 2021. Test conducted: 28 May 2020 to 21 April 2021*. Published 29 April 2021. Date accessed May 3, 2021. https://www.gov.uk/government/publications/weekly-statistics-for-nhs-test-and-trace-england-15-april-to-21-april-2021

Diggle, P. J. (2011). Estimating prevalence using an imperfect test. *Epidemiology Research International*, *2011*, 608719. https://doi.org/10.1155/2011/608719

Dinnes, J., Deeks, J. J., Berhane, S., Taylor, M., Adriano, A., Davenport, C., Dittrich, S., Emperador, D., Takwoingi, Y., Cunningham, J., Beese, S., Domen, J., Dretzke, J., Ferrante di Ruffano, L., Harris, I. M., Price, M. J., Taylor-Phillips, S., Hooft, L., Leeflang, M. M. G., … Van den Bruel, A. (2021). Rapid, point-of-care antigen and molecular-based tests for diagnosis of SARS-CoV-2 infection. *Cochrane Database of Systematic Reviews*, *3*, Art. No.: CD013705. https://doi.org/10.1002/14651858.CD013705.pub2

Doust, J. A., Bell, K. J. L., Leeflang, M. M. G., Dinnes, J., Lord, S. J., Mallett, S., van de Wijgert, J. H. H. M., Sandberg, S., Adeli, K., Deeks, J. J., Bossuyt, P. M., & Horvarth, A. R. (2021). Guidance for the design and reporting of studies evaluating the clinical performance of tests for present or past SARS-CoV-2 infection. *BMJ (Clinical Research Ed)*, *372*, n568. https://doi.org/10.1136/bmj.n568

Dwan, K., Gamble, C., Williamson, P. R., & Kirkham, J. J.; the Reporting Bias Group (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias—An updated review. *PLoS One*, *8*(7), e66844. https://doi.org/10.1371/journal.pone.0066844

Ehrmeyer, S. S., & Laessig, R. H. (2007). Point-of-care testing, medical error, and patient safety: A 2007 assessment. *Clinical Chemistry and Laboratory Medicine*, *45*(6), 766–773. https://doi.org/10.1515/CCLM.2007.164

European Parliament and Council. (1998). *Directive 98/79/EC of the European Parliament and of the Council of 27 October 1998 on in vitro diagnostic medical devices*. Date accessed May 1, 2021 https://eur-lex.europa.eu/legal-content/EN/TXT/? uri=CELEX:31998L0079

European Parliament and Council. (2017). *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*. Date accessed May 1, 2021 https://eur-lex.europa.eu/legal-content/EN/TXT/? uri=CELEX%3A32017R0746&qid=1620089932271

Ferguson, J., Dunn, S., Best, A., Mirza, J., Percival, B., Mayhew, M., Megram, O., Ashford, F., White, T., Moles-Garcia, E., Crawford, L., Plant, T., Bosworth, A., Kidd, M., Richter, A., Deeks, J., & McNally, A. (2021). Validation testing to determine the sensitivity of lateral flow testing for asymptomatic SARS-CoV-2 detection in low prevalence settings: Testing frequency and public health messaging is key. *PLoS Biology*, *19*(4), e3001216. https://doi.org/10.1371/journal.pbio.3001216

Ferrante di Ruffano, L., Davenport, C., Eisinga, A., Hyde, C., & Deeks, J. J. (2012). A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. *Journal of Clinical Epidemiology*, *65*(3), 282–287. https://doi.org/10.1016/j.jclinepi.2011.07.003

Ferrante di Ruffano, L., Dinnes, J., Sitch, A. J., Hyde, C., & Deeks, J. J. (2017). Test-treatment RCTs are susceptible to bias: A review of the methodological quality of randomized trials that evaluate diagnostic tests. *BMC Medical Research Methodology*, *17*(1), 35. https://doi.org/10.1186/s12874-016-0287-z

Ferrante di Ruffano, L., Hyde, C. J., McCaffery, K. J., Bossuyt, P. M., & Deeks, J. J. (2012). Assessing the value of diagnostic tests: A framework for designing and evaluating trials. *BMJ (Clinical Research Ed)*, *344*, e686. https://doi.org/10.1136/bmj.e686

FIND. (2021). *SARS-CoV-2 diagnostic use cases*. Date accessed April 18, 2024. https://www.finddx.org/covid-19/

Flower, B., Brown, J. C., Simmons, B., Moshe, M., Frise, R., Penn, R., Kugathasan, R., Petersen, C., Daunt, A., Ashby, D., Riley, S., Atchison, C. J., Taylor, G. P., Satkunarajah, S., Naar, L., Klaber, R., Badhan, A., Rosadas, C., Khan, M., … Cooke, G. S. (2020). Clinical and laboratory evaluation of SARS-CoV-2 lateral flow assays for use in a national COVID-19 seroprevalence survey. *Thorax*, *75*(12), 1082–1088. https://doi.org/10.1136/thoraxjnl-2020-215732

Food and Drug Administration (FDA). (2007, March). *Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests—Guidance for Industry and FDA Staff*. US Department of Health and Human Services. Date accessed May 6, 2021. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/statistical-guidance-reporting-results-studies-evaluating-diagnostic-tests-guidance-industry-and-fda

Food and Drug Administration (FDA). (2023, February). *In vitro diagnostics*. Date accessed April 18, 2024. https://www.fda.gov/medical-devices/products-and-medical-procedures/vitro-diagnostics

Fryback, D. G., & Thornbury, J. R. (1991). The efficacy of diagnostic imaging. *Medical Decision Making*, *11*(2), 88–94. https://doi.org/10.1177/0272989X9101100203

Glasziou, P., Altman, D. G., Bossuyt, P., Boutron, I., Clarke, M., Julious, S., Michie, S., Moher, D., & Wager, E. (2014). Reducing waste from incomplete or unusable reports of biomedical research. *Lancet (London, England)*, *383*(9913), 267–276. https://doi.org/10.1016/S0140-6736(13)62228-X

Glasziou, P., Irwig, L., & Deeks, J. J. (2008). When should a new test become the current reference standard? *Annals of Internal Medicine*, *149*(11), 816–822. https://doi.org/10.7326/0003-4819-149-11-200812020-00009

Goodacre, S., Irving, A., Wilson, R., Beever, D., & Challen, K. (2015). The PAndemic INfluenza Triage in the Emergency Department (PAINTED) pilot cohort study. *Health Technology Assessment (Winchester, England)*, *19*(3), v–69. https://doi.org/10.3310/hta19030

Gore, S. M., Bird, A. G., Burns, S. M., Goldberg, D. J., Ross, A. J., & Macgregor, J. (1995). Drug injection and HIV prevalence in inmates of Glenochil prison. *BMJ (Clinical Research Ed)*, *310*(6975), 293–296. https://doi.org/10.1136/bmj.310.6975.293

Gore, S. M., Bird, A. G., Cameron, S. O., Hutchinson, S. J., Burns, S. M., & Goldberg, D. J. (1999). Prevalence of hepatitis C carriage in Scottish prisons: WASH-C surveillance linked to self-reported risk behaviours. *QJM: An International Journal of Medicine*, *92*(1), 25–32. https://doi.org/10.1093/qjmed/92.1.25

Grijalva, C. G., Zhu, Y., Halasa, N. B., Kim, A., Rolfes, M. A., Steffens, A., Reed, C., Fry, A. M., & Talbot, H. (2020). High concordance between self-collected nasal swabs and saliva samples for detection of SARS-CoV-2. *Open Forum Infectious Diseases*, *7*(Suppl_1), S283. https://doi.org/10.1093/ofid/ofaa439.626

Gudbjartsson, D. F., Norddahl, G. L., Melsted, P., Gunnarsdottir, K., Holm, H., Eythorsson, E., Arnthorsson, A. O., Helgason, D., Bjarnadottir, K., Ingvarsson, R. F., Thorsteinsdottir, B., Kristjansdottir, S., Birgisdottir, K., Kristinsdottir, A. M., Sigurdsson, M. I., Arnadottir, G. A., Ivarsdottir, E. V., Andresdottir,

M., Jonsson, F., … Stefansson, K. (2020). Humoral immune response to SARS-CoV-2 in Iceland. *The New England Journal of Medicine*, *383*(18), 1724–1734. https://doi.org/10.1056/NEJMoa2026116

Gülmezoglu, A. M., Pang, T., Horton, R., & Dickersin, K. (2005). WHO facilitates international collaboration in setting standards for clinical trial registration. *Lancet (London, England)*, *365*(9474), 1829–1831. https://doi.org/10.1016/S0140-6736(05)66589-0

Haan, L., & Ferreira, A. (2007). *Extreme value theory: An introduction*. Springer.

Hadgu, A. (1999). Discrepant analysis: A biased and an unscientific method for estimating test sensitivity and specificity. *Journal of Clinical Epidemiology*, *52*(12), 1231–1237. https://doi.org/10.1016/s0895-4356(99)00101-8

Hall, V. J., Foulkes, S., Charlett, A., Atti, A., Monk, E. J. M., Simmons, R., Wellington, E., Cole, M. J., Saei, A., Oguti, B., Munro, K., Wallace, S., Kirwan, P. D., Shrotri, M., Vusirikala, A., Rokadiya, S., Kall, M., Zambon, M., Ramsay, M., … Hopkins, S. (2021). SARS-CoV-2 infection rates of antibody-positive compared with antibody-negative health-care workers in England: A large, multicentre, prospective cohort study (SIREN). *Lancet (London, England)*, *397*(10283), 1459–1469. https://doi.org/10.1016/S0140-6736(21)00675-9

Higgins, J. P. T., Savović, J., Page, M. J., Elbers, R. G., & Sterne, J. A. C. (2023). Chapter 8: Assessing risk of bias in a randomized trial. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane handbook for systematic reviews of interventions version 6.4 (updated August 2023)* (pp. 205–228). Cochrane. www.training.cochrane.org/handbook

Holtman, G. A., Berger, M. Y., Burger, H., Deeks, J. J., Donner-Banzhoff, N., Fanshawe, T. R., Koshiaris, C., Leeflang, M., Oke, J., Perera, R., Reitsma, J., & Van den Bruel, A. (2019). Development of practical recommendations for diagnostic accuracy studies in low-prevalence situations. *Journal of Clinical Epidemiology*, *114*, 38–48. https://doi.org/10.1016/j.jclinepi.2019.05.018

Horvath, A. R., Lord, S. J., StJohn, A., Sandberg, S., Cobbaert, C. M., Lorenz, S., Monaghan, P. J., Verhagen-Kamerbeek, W. D. J., Ebert, C., & Bossuyt, P. M. M.; Test Evaluation Working Group of the European Federation of Clinical Chemistry Laboratory Medicine (2014). From biomarkers to medical tests: The changing landscape of test evaluation. *Clinica Chimica Acta*, *427*, 49–57. https://doi.org/10.1016/j.cca.2013.09.018

Hung, I. F.-N., Cheng, V. C.-C., Li, X., Tam, A. R., Hung, D. L.-L., Chiu, K. H.-Y., Yip, C. C.-Y., Cai, J.-P., Ho, D. T.-Y., Wong, S.-C., Leung, S. S.-M., Chu, M.-Y., Tang, M. O.-Y., Chen, J. H.-K., Poon, R. W.-S., Fung, A. Y.-F., Zhang, R. R., Yan, E. Y.-W., Chen, L.-L., … Yuen, K.-Y. (2020). SARS-CoV-2 shedding and seroconversion among passengers quarantined after disembarking a cruise ship: A case series. *The Lancet Infectious Diseases*, *20*(9), 1051–1060. https://doi.org/10.1016/S1473-3099(20)30364-9

Hutchinson, S. J., Gore, S. M., Taylor, A., Goldberg, D. J., & Frischer, M. (2000). Extent and contributing factors of drug expenditure of injectors in Glasgow. Multi-site city-wide cross-sectional study. *The British Journal of Psychiatry*, *176*(2), 166–172. https://doi.org/10.1192/bjp.176.2.166

Innova Medical Group. (2020). *SARS-CoV-2 antigen rapid qualitative test. Instructions for use. Version A/02 2020–07-01*. Date accessed May 19, 2021. https://cdn.website-editor.net/6f54caea7c6f4adfba8399428f3c0b0c/files/uploaded/Innova-SARS-Cov-2-Antigen-test-IFU.pdf

Innova Medical Group. (2021). *Innova rapid antigen test as a public health screening tool*. Date accessed January 27, 2024. https://innovamedgroup.us/innova-rapid-antigen-test-as-a-public-health-screening-tool/

ISO (International Organisation for Standardization). (2019). *In vitro diagnostic medical devices—Clinical performance studies using specimens from human subjects—Good study practice*. ISO 20916:2019. Date accessed April 19, 2021. https://standards.iteh.ai/catalog/standards/sist/a4770d92-b282-49e5-b2e1-c719ca76aa1b/iso-20916-2019

Iyer, A. S., Jones, F. K., Nodoushani, A., Kelly, M., Becker, M., Slater, D., Mills, R., Teng, E., Kamruzzaman, M., Garcia-Beltran, W. F., Astudillo, M., Yang, D., Miller, T. E., Oliver, E., Fischinger, S., Atyeo, C., Iafrate, A. J., Calderwood, S. B., Lauer, S. A., … Charles, R. C. (2020). Persistence and decay of human antibody responses to the receptor binding domain of SARS-CoV-2 spike protein in COVID-19 patients. *Sci Immunol*, *5*(52), eabe0367. https://doi.org/10.1126/sciimmunol.abe0367

Johnson, R. (2008). Assessment of bias with emphasis on method comparison. *The Clinical Biochemist Reviews*, *29*(Suppl 1), S37–S42. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2556581/

Korevaar, D. A., Gopalakrishna, G., Cohen, J. F., & Bossuyt, P. M. (2019). Targeted test evaluation: A framework for designing diagnostic accuracy studies with clear study hypotheses. *Diagnostic and Prognostic Research*, *3*(1), 22. https://doi.org/10.1186/s41512-019-0069-2

Larremore, D. B., Wilder, B., Lester, E., Shehata, S., Burke, J. M., Hay, J. A., Tambe, M., Mina, M. J., & Parker, R. (2021). Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening. *Science Advances*, *7*(1), eabd5393. https://doi.org/10.1126/sciadv.abd5393

Lijmer, J. G., Leeflang, M., & Bossuyt, P. M. (2009). Proposals for a phased evaluation of medical tests. *Medical Decision Making*, *29*(5), E13–E21. https://doi.org/10.1177/0272989X09336144

Lijmer, J. G., Mol, B. W., Heisterkamp, S., Bonsel, G. J., Prins, M. H., van der Meulen, J. H., & Bossuyt, P. M. (1999). Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*, *282*(11), 1061–1066. https://doi.org/10.1001/jama.282.11.1061

Lord, S. J., Staub, L. P., Bossuyt, P. M. M., & Irwig, L. M. (2011). Target practice: Choosing target conditions for test accuracy studies that are relevant to clinical practice. *BMJ (Clinical Research Ed)*, *343*, d4684. https://doi.org/10.1136/bmj.d4684

Love, N. K., Ready, D. R., Turner, C., Verlander, N. Q., French, C. E., Martin, A. F., Sorensen, T. B., Metelmann, S., Denford, S., Rubin, G. J., Yardley, L., Amlôt, R., Hopkins, S., & Oliver, I. (2022). Daily use of lateral flow devices by contacts of confirmed COVID-19 cases to enable exemption from isolation compared with standard self-isolation to reduce onward transmission of SARS-CoV-2 in England: A randomised, controlled, non-inferiority trial. *The Lancet Respiratory Medicine*, *10*(11), 1074–1085. https://doi.org/10.1016/S2213-2600(22)00267-3

Machalaba, C. C., Loh, E. H., Daszak, P., & Karesh, W. B. (2015). Emerging diseases from animals. *State of the World*, *2015*, 105–116. https://doi.org/10.5822/978-1-61091-611-0_8

Mandavilli, A. (2020). The White House relied on a rapid test, but used it in a way it was not intended. *The New York Times*. Date accessed May 18, 2021. https://www.nytimes.com/2020/10/02/us/elections/the-white-house-relied-on-a-rapid-test-but-used-it-in-a-way-it-was-not-intended.html

McInnes, M. D. F., Moher, D., Thombs, B. D., McGrath, T. A., Bossuyt, P. M.; and the PRISMA-DTA Group; Clifford, T., Cohen, J. F., Deeks, J. J., Gatsonis, C., Hooft, L., Hunt, H. A., Hyde, C. J., Korevaar, D. A., Leeflang, M. M. G., Macaskill, P., Reitsma, J. B., Rodin, R., Rutjes, A. W. S., … Willis, B. H. (2018). Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: The PRISMA-DTA statement. *JAMA*, *319*(4), 388–396. https://doi.org/10.1001/jama.2017.19163

Medicines and Healthcare products Regulatory Agency. (2018). *Corporate report: MHRA Corporate Plan 2018 to 2023*. Date accessed May 1, 2021. https://www.gov.uk/government/publications/mhra-corporate-plan-2018-to-2023

Medicines and Healthcare products Regulatory Agency. (2020a, May 29). *New story: Action taken to halt sales of fingerprick coronavirus (Covid-19) antibody testing kits*. Date accessed January 27, 2024. https://www.gov.uk/government/news/action-taken-to-halt-sales-of-fingerprick-coronavirus-covid-19-antibody-testing-kits

Medicines and Healthcare products Regulatory Agency. (2020b). *Target product profile: In Vitro Diagnostic (IVD) self-tests for the detection of SARS-CoV-2 in people without symptoms*. Date accessed June 9, 2021. https://www.gov.uk/government/publications/how-tests-and-testing-kits-for-coronavirus-covid-19-work.

Medicines and Healthcare products Regulatory Agency. (2020c). *Decision: Medical devices given exceptional use authorisations during the Covid-19 pandemic*. Date accessed May 1, 2021. https://www.gov.uk/government/publications/medical-devices-given-exceptional-use-authorisations-during-the-covid-19-pandemic

Medicines and Healthcare products Regulatory Agency. (2023, July). *Guidance on the regulation of In Vitro Diagnostic medical devices in Great Britain*. Guidance on the regulation of IVD medical devices in GB (publishing.service.gov.uk). Date accessed January 31, 2024. https://assets.publishing.service.gov.uk/media/64bfd78a1e10bf000d17ce1e/Guidance_on_the_regulation_of_IVD_medical_devices_in_GB.pdf

Mitjà, O., Corbacho-Monné, M., Ubals, M., Alemany, A., Suñer, C., Tebé, C., Tobias, A., Peñafiel, J., Ballana, E., Pérez, C. A., Admella, P., Riera-Martí, N., Laporte, P., Mitjà, J., Clua, M., Bertran, L., Sarquella, M., Gavilán, S., Ara, J., … Clotet, B. (2021). A cluster-randomized trial of hydroxychloroquine for prevention of COVID-19. *The New England Journal of Medicine*, *384*(5), 417–427. https://doi.org/10.1056/NEJMoa2021801

Mizumoto, K., Kagaya, K., Zarebski, A., & Chowell, G. (2020). Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the diamond princess cruise ship, Yokohama, Japan, 2020. *Euro Surveillance*, *25*(10), 2000180. https://doi.org/10.2807/1560- 7917.ES.2020.25.10.2000180

Moshe, M., Daunt, A., Flower, B., Simmons, B., Brown, J. C., Frise, R., Penn, R., Kugathasan, R., Petersen, C., Stockmann, H., Ashby, D., Riley, S., Atchison, C., Taylor, G. P., Satkunarajah, S., Naar, L., Klaber, R., Badhan, A., Rosadas, C., … Barclay, W. S. (2021). SARS-CoV-2 lateral flow assays for possible use in national COVID-19 seroprevalence surveys (react 2): Diagnostic accuracy study. *BMJ (Clinical Research Ed)*, *372*, n423. https://doi.org/10.1136/bmj.n423

Mulchandani, R., Jones, H. E., Taylor-Phillips, S., Shute, J., Perry, K., Jamarani, S., Brooks, T., Charlett, A., Hickman, M., Oliver, I., Kaptoge, S., Danesh, J., Di Angelantonio, E., Ades, A. E., & Wyllie, D. H.; on behalf of the EDSAB-HOME and COMPARE Investigators. (2020). Accuracy of UK Rapid Test Consortium (UK-RTC) "AbC-19 rapid test" for detection of previous SARS-CoV-2 infection in key workers: Test accuracy study. *BMJ (Clinical Research Ed)*, *371*, m4262 https://doi.org/10.1136/bmj.m4262

Naaktgeboren, C. A., Bertens, L. C. M., van Smeden, M., de Groot, J. A. H., Moons, K. G. M., & Reitsma, J. B. (2013). Value of composite reference standards in diagnostic research. *BMJ (Clinical Research Ed)*, *347*(oct25 2), f5605. https://doi.org/10.1136/bmj.f5605

Naudet, F., Patel, C. J., DeVito, N. J., Le Goff, G., Cristea, I. A., Braillon, A., & Hoffmann, S. (2024). Improving the transparency and reliability of observational studies through registration. *BMJ (Clinical Research Ed)*, *384*, e076123. https://doi.org/10.1136/bmj-2023-076123

NHS Test and Trace. (2020, December 15). *Covid-19 National Testing Programme: Schools & Colleges Handbook*. Date accessed May 19, 2021. https://schoolsweek.co.uk/wp-content/uploads/2020/12/Schools_Colleges_Testing-Handbook_version-3.3-Copy.pdf

Obuchowski, N. A. (1998). Sample size calculations in studies of test accuracy. *Statistical Methods in Medical Research*, *7*(4), 371–392. https://doi.org/10.1177/096228029800700405

Office for National Statistics. (2020). *Coronavirus (Covid-19) Infection Survey, UK: 21 August 2020*. Date accessed May 19, 2021. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveypilot/englandandwales21august2020

Our World in Data. (2021). *Australia: Coronavirus Pandemic Country Profile*. Date accessed May 1, 2021. https://ourworldindata.org/coronavirus/country/australia

Peto, T.; UK Covid-19 Lateral Flow Oversight Team. (2021). COVID-19: Rapid antigen detection for SARS-CoV-2 by lateral flow assay: A national systematic evaluation for mass-testing. *EClinicalMedicine*, *36*, 100924. https://doi.org/10.1016/j.eclinm.2021.100924

Pickering, S., Batra, R., Merrick, B., Snell, L. B., Nebbia, G., Douthwaite, S., Reid, F., Patel, A., Kia Ik, M. T., Patel, B., Charalampous, T., Alcolea-Medina, A., Lista, M. J., Cliff, P. R., Cunningham, E., Mullen, J., Doores, K. J., Edgeworth, J. D., Malim, M. H., … Galão, R. P. (2021). Comparative performance of SARS-CoV-2 lateral flow antigen tests and association with detection of infectious virus in clinical specimens: A single-centre laboratory evaluation study. *The Lancet Microbe*, *2*(9), e461–e471. https://doi.org/10.1016/S2666-5247(21)00143-9. Epub 2021 Jun 30. Erratum in: *The Lancet Microbe*, 2021, *2*(9), e426.

Post, N., Eddy, D., Huntley, C., van Schalkwyk, M. C. I., Shrotri, M., Leeman, D., Rigby, S., Williams, S. V., Bermingham, W. H., Kellam, P., Maher, J., Shields, A. M., Amirthalingam, G., Peacock, S. J., & Ismail, S. A. (2020). Antibody response to SARS-CoV-2 infection in humans: A systematic review. *PLoS One*, *15*(12), e0244126. https://doi.org/10.1371/journal.pone.0244126

Pray, I. W., Ford, L., Cole, D., Lee, C., Bigouette, J. P., Abedi, G. R., Bushman, D., Delahoy, M. J., Currie, D., Cherney, B., Kirby, M., Fajardo, G., Caudill, M., Langolf, K., Kahrs, J., Kelly, P., Pitts, C., Lim, A., Aulik, N., … Killerby, M. (2021). Performance of an antigen-based test for asymptomatic and symptomatic SARS-CoV-2 testing at two university campuses—Wisconsin, September-October 2020. *MMWR: Morbidity and Mortality Weekly Report*, *69*(5152), 1642–1647. https://doi.org/10.15585/mmwr.mm695152a3

Public Health England. (2020, July). *Evaluation of sensitivity and specificity of four commercially available SARS-CoV-2 antibody immunoassays*. Date accessed April 12, 2021. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/898437/Evaluation__of_sensitivity_and_specificity_of_4_commercially_available_SARS-CoV-2_antibody_immunoassays.pdf

Quilty, B. J., Clifford, S., Hellewell, J., Russell, T. W., Kucharski, A. J., Flasche, S., & Edmunds, W. J.; Centre for the Mathematical Modelling of Infectious Diseases Covid-19 working group (2021). Quarantine and testing strategies in contact tracing for SARS-CoV-2: A modelling study. *Lancet Public Health*, *6*(3), e175–e183. https://doi.org/10.1016/S2468-2667(20)30308-X

Ramdas, K., Darzi, A., & Jain, S. (2020). 'Test, re-test, re-test': Using inaccurate tests to greatly increase the accuracy of COVID-19 testing. *Nature Medicine*, *26*(6), 810–811. https://doi.org/10.1038/s41591-020-0891-7

Ransohoff, D. F., & Feinstein, A. R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *The New England Journal of Medicine*, *299*(17), 926–930. https://doi.org/10.1056/NEJM197810262991705

Revollo, B., Blanco, I., Soler, P., Toro, J., Izquierdo-Useros, N., Puig, J., Puig, X., Navarro-Pérez, V., Casañ, C., Ruiz, L., Perez-Zsolt, D., Videla, S., Clotet, B., & Llibre, J. M. (2021). Same-day SARS-CoV-2 antigen test screening in an indoor mass-gathering live music event: A randomised controlled trial. *The Lancet Infectious Diseases*, *21*(10), 1365–1372. https://doi.org/10.1016/S1473-3099(21)00268-1

REACT Study Investigators; Riley, S., Ainslie, K. E., Eales, O., Walters, C. E., Wang, H., Atchison, C., Diggle, P. J., Ashby, D., Donnelly, C. A., Cooke, G., Barclay, W., Ward, H., Darzi, A., & Elliott, P. (2020). *Transient dynamics of SARS-CoV-2 as England exited national lockdown*. medRxiv 2020.08.05.20169078. https://doi.org/10.1101/2020.08.05.20169078, preprint: not peer reviewed.

Robertson, L. J., Moore, J. S., Blighe, K., Ng, K. Y., Quinn, N., Jennings, F., Warnock, G., Sharpe, P., Clarke, M., Maguire, K., Rainey, S., Price, R., Burns, W. P., Kowalczyk, A. M., Awuah, A., McNamee, S. E., Wallace, G. E., Hunter, D., Segar, S., … Moore, T. (2021). Evaluation of the IgG antibody response to SARS CoV-2 infection and performance of a lateral flow immunoassay: cross-sectional and longitudinal analysis over 11 months. *BMJ Open*, *11*(6), e048142. https://doi.org/10.1136/bmjopen-2020-048142

Royal Society of Medicine. (2020). *Covid-19 series: Challenges of testing—Episode 38. Professor Sir John Bell [Interview]*. Date accessed January 27, 2024. https://www.youtube.com/watch?v=WYpiFJJPP4s&list=PL1yY3WBVbNWAUNRjD5SUeC9R-G9vT8Gvq&index=59

Royal Statistical Society. (2020). *Covid-19 Taskforce*. Date accessed April 24, 2024. https://rss.org.uk/policy-campaigns/policy-groups/covid-19-task-force/

Royal Statistical Society. (2021, March 5). *Covid-19 Taskforce. Statement on the need for transparency about information on secondary pupils' update of Lateral Flow Testing, PCR-corroboration, cycle-threshold-values (proxy for viral laod) and genomic analyses*. Date accessed May 19, 2021. https://rss.org.uk/RSS/media/File-library/News/2021/RSS-statement-on-surveillance-in-schools-5-March-2021.pdf

Rutjes, A. W., Reitsma, J. B., Vandenbroucke, J. P., Glas, A. S., & Bossuyt, P. M. (2005). Case-control and two-gate designs in diagnostic accuracy studies. *Clinical Chemistry*, *51*(8), 1335–1341. https://doi.org/10.1373/clinchem.2005.048595

Salameh, J. P., Bossuyt, P. M., McGrath, T. A., Thombs, B. D., Hyde, C. J., Macaskill, P., Deeks, J. J., Leeflang, M., Korevaar, D. A., Whiting, P., Takwoingi, Y., Reitsma, J. B., Cohen, J. F., Frank, R. A., Hunt, H. A., Hooft, L., Rutjes, A. W. S., Willis, B. H., Gatsonis, C., … McInnes, M. D. F. (2020). Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): Explanation, elaboration, and checklist. *BMJ (Clinical Research Ed)*, *370*, m2632. https://doi.org/10.1136/bmj.m2632

Schuit, E., Veldhuijzen, I. K., Venekamp, R. P., van den Bijllaardt, W., Pas, S. D., Lodder, E. B., Molenkamp, R., GeurtsvanKessel, C. H., Velzing, J., Huisman, R. C., Brouwer, L., Boelsums, T., Sips, G. J., Benschop, K. S. M., Hooft, L., van de Wijgert, J. H. H. M., van den Hof, S., & Moons, K. G. M. (2021). Diagnostic accuracy of rapid antigen tests in asymptomatic and presymptomatic close contacts of individuals with confirmed SARS-CoV-2 infection: Cross sectional study. *medRxiv 374*, n1676. https://doi.org/10.1136/bmj.n1676

Science and Technology Committee. (2014). *Science and Technology Committee—4 Communicating the risks and benefits of health screening*. National Health Screening. Point 52. Date accessed May 19, 2021. https://publications.parliament.uk/pa/cm201415/cmselect/cmsctech/244/24407.htm#a13

Scientific Advisory Group for Emergencies. (2020). *Minutes Meeting 73, 17th December*. Date accessed May 19, 2021. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/952613/s0989-covid-19-sage-73-minutes-171220.pdf

Shinkins, B., Thompson, M., Mallett, S., & Perera, R. (2013). Diagnostic accuracy studies: How to report and analyse inconclusive test results. *BMJ (Clinical Research Ed)*, *346*, f2778. https://doi.org/10.1136/bmj.f2778

Shuren, J., & Stenzel, T. (2020). COVID-19 molecular diagnostic testing—Lessons learned. *The New England Journal of Medicine*, *383*(17), e97. https://doi.org/10.1056/NEJMp2023830

Shuren, J., & Stenzel, T. (2021). The FDA's experience with COVID-19 antibody tests. *The New England Journal of Medicine*, *384*(7), 592–594. https://doi.org/10.1056/NEJMp2033687

Simes, R. J. (1986). Publication bias: The case for an international registry of clinical trials. *Journal of Clinical Oncology*, *4*(10), 1529–1541. https://doi.org/10.1200/JCO.1986.4.10.1529

Singanayagam, A., Patel, M., Charlett, A., Lopez, B. J., Vanessa, S., Joanna, E., Shamez, L., Maria, Z., & Robin, G. (2020). Duration of infectiousness and correlation with RT-PCR cycle threshold values in cases of COVID-19, England, January to May 2020. *Euro Surveillance*, *25*(32):2001483. https://doi.org/10.2807/1560-7917.ES.2020.25.32.2001483

Spiegelhalter, D., & Masters, A. (2021). *COVID by numbers: Making sense of the pandemic with data*. Pelican.

Taichman, D. B., Sahni, P., Pinborg, A., Peiperl, L., Laine, C., James, A., Hong, S. T., Haileamlak, A., Gollogly, L., Godlee, F., Frizelle, F. A., Florenzano, F., Drazen, J. M., Bauchner, H., Baethge, C., & Backus, J. (2017). Data sharing statements for clinical trials: A requirement of the international committee of medical journal editors. *Ethiopian Journal of Health Sciences*, *27*(4), 315–318. https://doi.org/10.4314/ejhs.v27i4.2

Takwoingi, Y. (2016). *Meta-analytic approaches for summarising and comparing the accuracy of medical tests*. University of Birmingham. Ph.D. Date accessed May 19, 2021. https://etheses.bham.ac.uk/id/eprint/6759/

Takwoingi, Y., Leeflang, M. M., & Deeks, J. J. (2013). Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Annals of Internal Medicine*, *158*(7), 544–554. https://doi.org/10.7326/0003-4819-158-7-201304020-00006

Tholen, D., Linnet, K., Kondratovich, M., Armbruster, D., Garrett, P. E., Jones, R., Kroll, M. H., Lequin, R., Pankratz, T., Scassellati, G. A., Schimmel, H., & Tsai, J. (2004). *Protocols for determination of limits of detection and limits of quantitation; approved guidelines. NCCLS document EP17-A* (Vol. *24*, pp. 1–39). National Committee of Clinical and Laboratory Standards.

Thombs, B. D., & Rice, D. B. (2016). Sample sizes and precision of estimates of sensitivity and specificity from primary studies on the diagnostic accuracy of depression screening tools: A survey of recently published studies. *International Journal of Methods in Psychiatric Research*, *25*(2), 145–152. https://doi.org/10.1002/mpr.1504

University of Bristol. (2011). *QUADAS-2*. Date accessed April 18, 2024. https://www.bristol.ac.uk/population-health-sciences/projects/quadas/quadas-2/

University of Liverpool. (2020). Liverpool Covid-19 community testing pilot. Interim evaluation report. Date accessed January 27, 2024. https://www.liverpool.ac.uk/media/livacuk/research/Liverpool, Community, Testing, Pilot, Interim, Evaluation.pdf

van Seventer, J. M., & Hochberg, N. S. (2017). Principles of infectious diseases: Transmission, diagnosis, prevention, and control. *International Encyclopedia of Public Health*, 22–39. https://doi.org/10.1016/B978-0-12-803678-5.00516-6

Ward, H., Cooke, G. S., Atchison, C., Whitaker, M., Elliott, J., Moshe, M., Brown, J. C., Flower, B., Daunt, A., Ainslie, K., Ashby, D., Donnelly, C. A., Riley, S., Darzi, A., Barclay, W., & Elliott, P. (2021a) Prevalence of antibody positivity to SARS-CoV-2 following the first peak of infection in England: Serial cross-sectional studies of 365,000 adults. *The Lancet Regional Health – Europe*. *18*, 100098. https://doi.org/ 10.1016/j.lanepe.2021.100098

Ward, H., Cooke, G., Whitaker, M., Redd, R., Eales, O., Brown, J. C., Collet, K., Cooper, E., Daunt, A., Jones, K., Moshe, M., Willicombe, M., Day, S., Atchison, C., Darzi, A., Donnelly, C. A., Riley, S., Ashby, D., Barclay, W. S., & Elliott, P. (2021b). *REACT-2 Round 5: increasing prevalence of SARS-CoV-2 antibodies demonstrate impact of the second wave and of vaccine roll-out in England*. medRxiv 2021.02.26.21252512. https://doi.org/10.1101/2021.02.26.21252512, preprint: not peer reviewed.

Weijer, C., Hemming, K., Phillips Hey, S, & Fernandez Lynch, H. (2021). Reopening schools safely in the face of COVID-19: Can cluster randomized trials help? *Clinical Trials*, *18*(3), 371–376. https://doi.org/10.1177/1740774520984860

Welch, H. G., & Black, W. C. (1997). Evaluating randomized trials of screening. *Journal of General internal Medicine*, *12*(2), 118–124. https://doi.org/10.1046/j.1525-1497.1997.00017.x

White, S. R., Hutchinson, S. J., Taylor, A., & Bird, S. M. (2015). Modeling the initiation of others into injection drug use, using data from 2,500 injectors surveyed in Scotland during 2008–2009. *American Journal of Epidemiology*, *181*(10), 771–780. https://doi.org/10.1093/aje/kwu345

Whiting, P. F., Rutjes, A. W., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., Leeflang, M. M., Sterne, J. A., & Bossuyt, P. M. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine*, *155*(8), 529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

World Health Organization. (2020, March 20). *Global surveillance for Covid-19 caused by human infection with Covid-19 virus: Interim guidance*. Date accessed May 21, 2021. https://www.who.int/publications/i/item/who-2019-nCoV-surveillanceguidance-2020.8

World Health Organization. (2021, March 30). *WHO-convened global study of origins of SARS-CoV-2: China Part*. Date accessed April 18, 2024. https://www.who.int/publications/i/item/who-convened-global-study-of-origins-of-sars-cov-2-china-part

World Health Organization. (2024). *Disease Outbreak News (DONs)*. Date accessed April 18, 2024. https://www.who.int/emergencies/disease-outbreak-news

Yirrell, D. L., Robertson, P., Goldberg, D. J., McMenamin, J., Cameron, S., & Leigh Brown, A. J. (1997). Molecular investigation into outbreak of HIV in a Scottish prison. *BMJ (Clinical Research Ed)*, *314*(7092), 1446–1450. https://doi.org/10.1136/bmj.314.7092.1446

Yordanov, Y., Dechartres, A., Atal, I., Tran, V. T., Boutron, I., Crequit, P., & Ravaud, P. (2018). Avoidable waste of research related to outcome planning and reporting in clinical trials. *BMC Medicine*, *16*(1), 87. https://doi.org/10.1186/s12916-018-1083-x

Young, B. C., Eyre, D. W., Kendrick, S., White, C., Smith, S., Beveridge, G., Nonnenmacher, T., Ichofu, F., Hillier, J., Oakley, S., Diamond, I., Rourke, E., Dawe, F., Day, I., Davies, L., Staite, P., Lacey, A., McCrae, J., Jones, F., … Peto, T. E. A. (2021). Daily testing for contacts of individuals with SARS-CoV-2 infection and attendance and SARS-CoV-2 transmission in English secondary schools and colleges: An open-label, cluster-randomised trial. *Lancet (London, England)*, *398*(10307), 1217–1229. https://doi.org/10.1016/S0140-6736(21)01908-5