ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

Royal Statistical Society Diagnostic Tests Working Group Report

JUNE 2021

Abbreviations

AIDS	Acquired immunodeficiency syndrome
ARC	Applied Research Collaborative
A&E	Accident and Emergency
BRC	Biomedical Research Centre
BSE	Bovine spongiform encephalopathy
CDC	Centers for Disease Control and Prevention
CE	Conformité Européene
CI	Confidence interval
CLIA	Chemiluminescent immunoassay
CLSI	Clinical and Laboratory Standards Institute
COVID-19	Coronavirus disease 2019
Ct	Cycle threshold
CV	Coefficient of variation
DAA	Directly acting antiviral
DHSC	Department for Health and Social Care
DNA	Deoxyribonucleic acid
ELISA	Enzyme-linked immunosorbent assay
EU	European Union
EUA	Emergency use approval
FDA	Food and Drug Administration
FIND	Foundation for Innovative New Diagnostics
FN	False negative
FP	False positive
HAART	Highly active antiretroviral therapies
HAV	Hepatitis A virus
HBV	Hepatitis B virus
HCV	Hepatitis C virus
HCW	Healthcare worker
HIV	Human immunodeficiency virus
HMG	Her Majesty's Government
IFU	Instructions for use
Ig	Immunoglobulin
IGRA	Interferon gamma release assays
ISO	International Organisation for Standards
IVD	In vitro diagnostic
LFT	Lateral flow test
LoB	Limit of blank
LoD	Limit of detection
LoQ	Limit of quantification
LTBI	Latent tuberculosis infection
MHRA	Medical and Healthcare products Regulatory Agency

MRC	Medical Research Council
MTB/RIF	Mycobacterium tuberculosis and resistance to rifampicin
NHS	National Health Service
NIHR	National Institute for Health Research
ONS	Office of National Statistics
NPV	Negative predictive value
PCR	Polymerase chain reaction
PHE	Public Health England
POCT	Point of care test
PPV	Positive predictive value
PRISMA-DTA	Preferred reporting items for systematic reviews and meta-analyses of diagnostic test accuracy studies
QUADAS-2	Quality assessment of diagnostic accuracy studies
RCV	Reference change value
RDT	Rapid diagnostic test
REACT	Real-time assessment of community transmission
RNA	Ribonucleic acid
RSS	Royal Statistical Society
RT-LAMP	Reverse transcription loop-mediated isothermal amplification
RT-PCR	Reverse transcription polymerase chain reaction
SAGE	Scientific Advisory Group for Emergencies
SARS	Severe acute respiratory syndrome
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SD	Standard deviation
STARD	Standards for reporting diagnostic accuracy
TB	Tuberculosis
TN	True negative
TP	True positive
TPP	Target product profile
TST	Tuberculin skin test
UAT	Unlinked anonymous testing
UK	United Kingdom
USA	United States of America
vCJD	Variant Creutzfeldt–Jakob disease
VTM	Viral transport media
WHO	World Health Organisation

Table of Contents

Executive Summary

Context

In early January 2020, a new coronavirus, SARS-CoV-2 causing COVID-19, was identified and had already begun its rapid global spread. The UK began its first lockdown on 23rd March 2020. Since then, science and scientists have moved at a great pace to combat this pandemic infection. Statistics has been playing many key roles: in infectious disease modelling, randomised controlled trials, infection control, licensing, surveillance and test-evaluation.

Key areas of statistical input included modelling to predict the spread under various assumptions, surveillance studies to monitor regularly (nationally and regionally) changes in the prevalence of current or previous infection, development of strategies for testing, tracing and isolation, trials of treatments in a variety of settings, and of vaccines. Each of these has multiple statistical issues, and so the Royal Statistical Society's (RSS) Council set up a COVID-19 Task Force with a key aim of ensuring that the RSS could contribute its collective expertise to the UK's national and devolved governments and public bodies, on statistical issues during the COVID-19 pandemic.

Because the pandemic developed so quickly, there was enormous pressure for rapidly available solutions, but at the risk of inadequate evaluation. The RSS has been particularly concerned that many new diagnostic tests for SARS-CoV-2 antigen or antibodies 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 wider background of concern about standards applied to the evaluation of *in vitro* diagnostic tests, there was a need for clear statistical thinking on the principles of diagnostic testing in general, and their application in a pandemic in particular.

This review has been undertaken by the RSS at this time as the SARS-CoV-2 pandemic has provided a microcosmic insight of the inadequate state of current processes for evaluating and regulating medical tests. Whilst directly motivated by the pandemic, the findings apply more broadly to *in vitro* diagnostics, and in part to all diagnostics.

Terms of reference

The key aim is to review the statistical evidence needed to assure the performance of new tests, for patients, decision-makers and regulators, with particular reference to *in vitro* diagnostics (IVDs) for infectious diseases.

Considerations include:

- 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.

Section 1 provides a background to infectious diseases and key terminology, and Section 2 outlines key concepts in diagnostic testing, with a more detailed exposition of statistical estimands in Section 3. Section 4 addresses considerations of good study design for evaluation of diagnostic tests. Section 5 addresses the SARS-CoV-2 pandemic specifically, and the lessons to be drawn. Section 6 deals with the implications for regulation of diagnostic test, and Section 7 the information that should be in the public domain.

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 **multisite 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 centralised processes for **ethical** and **study-protocol approvals**.

Recommendations: Regulation matters

1) **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.

2) 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**.

3) 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.

4) 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 liberalised behaviour to deprivation of liberty, should be considered.

5) Evaluation of the **impact of tests** should ensure that both **intended** and **unintended consequences** are considered. Some consequences will not be evaluable before test implementation, so that **post-marketing surveillance** for a new intended use requires ongoing assessment.

6) 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.

Recommendations: Transparency matters

1) **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**.

2) **Expert peer review** of study protocols and final reports by **subject-matter** (eg, clinical, public health, laboratory) and **methodology experts** is recommended.

3) **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.

4) **Post hoc analyses** should be limited; and clearly identified as **exploratory**.

5) **Study reports** and results should be made available **publicly** in a **timely manner**.

6) Field and clinical evaluation studies require **ethical approval** and **fully informed consent** as outlined in the **Good Clinical Practice Guidelines**.

Section 1: 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 on public health systems around the world.

Infectious organisms (pathogens) include viruses (eg, Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2)), bacteria (eg, *Mycobacterium tuberculosis*), parasites (eg, *Plasmodium* species), and fungi (eg, *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 (https://www.who.int/emergencies/disease-outbreak-news). Notable recent outbreaks include SARS (2002-2003), swine flu (2009-2011), Ebola (2014−2016), Zika (2015−2016), dengue (2016), plague (2017), and COVID-19 (2019−present). 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). Given the potential for significant mortality or morbidity, safe and effective vaccines to provide immunity against life-threatening infectious diseases are a critical preventive intervention.

Some infectious diseases are zoonotic, ie, 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 AD), the Black Death (Europe, 1347), yellow fever (South America, sixteenth century), and the Spanish flu (1918) (Machalaba *et al*, 2015). The SARS-CoV-2 pandemic appears to have zoonotic origins.

Transmissibility of infectious diseases combined with age-related susceptibility or progression requires those infected to take precautions to prevent onward transmission to others; and that all of us modify our behaviours to reduce our 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 this section outlines some necessary key biological concepts in 1.1 and 1.2.

1.1 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 immunisation 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 life-long 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 recognised by antibodies that may have been produced in response to infection with a previous strain of the virus.

1.2 Stages of infectious diseases

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, 2020). Figure 1.1 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 (eg, SARS-Cov-2), fomite or indirectly (eg, 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 *et al*, 2017). The interval between the time of exposure and onset of disease symptoms (if symptoms appear) is known as the **incubation phase.** This subclinical or asymptomatic phase can range from a few days (eg, SARS-CoV-2 up to 14 days) to several years (eg, 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 (eg, latent tuberculosis), or act as **carriers** able to transmit infection to others (eg, 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 et al, 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, for defining the intended use of a test, and for 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.

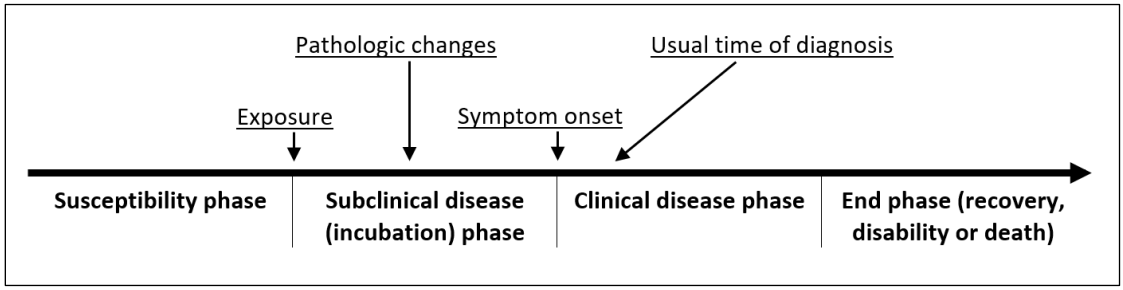


Figure 1.1 Natural history of disease timeline

Adapted from Figure 1.18 in Centers for Disease Control and Prevention (2006) Principles of Epidemiology in Public Health Practice, Third Edition. Available at: https://www.cdc.gov/csels/dsepd/ss1978/lesson1/section9.html (last accessed: May 2021).

Influenza and hepatitis remind us 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 United Kingdom's (UK) blood supply had to be protected despite there being no test in blood for the abnormal prion protein which causes vCJD, see Appendix 1**.**

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 have raised global challenges for clinical medicine, public health, our economies and everyday life. Development of new diagnostic tests is part of the necessary response. In this report we describe key issues in test evaluation of in vitro diagnostics for infectious disease and investigate how they have been addressed in the evaluation of new tests for SARS-CoV-2.

Section 2: Diagnostic tests

Diagnostic tests can indicate the presence or absence of infection or a surrogate marker of infection; or detect evidence of previous infection (eg, 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.

We focus here on *in vitro* diagnostic tests as these are the most common type of test for diagnosis of infectious diseases. However, many of the issues we raise are generic and will apply to other test types, such as imaging.

This section gives biological detail on types of in vitro diagnostic tests in 2.1, and possible settings for their use in 2.2. After outlining the challenges in 2.3, a framework is set in 2.4 for the key characteristics when evaluating a test.

2.1 Types of *in vitro* diagnostic tests for infectious diseases

In vitro diagnostics (IVDs) 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 (https://www.fda.gov/medical-devices/products-and-medical-procedures/vitro-diagnostics). See Box 1 for descriptions of different types of IVDs.

Box 1: Types of in vitro diagnostic tests for infectious disease

Microscopy

Microscopy enables visualisation of a pathogen by using a microscope to examine a specimen (eg, 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, eg, Giemsa stained thick or thin blood smear for detecting malaria parasites. Wet mounts (ie, 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 an artificial nutrient medium 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 in 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 time to diagnosis and access to testing, eg, 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 or absence 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, eg, 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.

2.2 Setting of testing

Laboratory testing can be performed in different settings, from centralised laboratories to self-testing in homes. Laboratory tests that are performed at the point of care, for example, at the bedside or in a clinic, rather than in a laboratory are often referred to as **point-of-care tests** (POCTs). Ehrmeyer (Ehrmeyer *et al*, 2007) defined POC 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 care-givers 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 be revolutionise 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. RDTs may be POCTs or laboratory-based tests.

2.3 Challenges in testing for infectious diseases

Most infectious diseases are caused by viruses, bacteria, or parasites. In Appendix 1, **pandemic and seasonal influenza** (viral), **hepatitis** (viral), **HIV** (virus), **vCJD** (abnormal prion protein), **tuberculosis** (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.

2.4 General principles

Principles applicable to the examples in Appendix 1 and other infectious diseases more generally are specification of:

- Target population (eg, adults, children, key risk groups)
- Setting (eg, low, moderate or high transmission setting, local conditions)
- Target condition (eg, stage of disease, pathogen identification, drug resistance)
- Reference standard (eg, single test or composed of multiple pieces of information, technical expertise required)
- Index test characteristics (eg, test format, specimen type, technical expertise)

These issues are important considerations when designing test evaluation studies and are addressed in Section 4.

Section 3: Statistical parameters required for reporting

Test evaluation from bench to bedside is multifarious. Horvath (Horvath *et al*, 2014) identified five key components of test evaluation: 1) **analytical performance**, 2) **clinical performance**, 3) **clinical effectiveness**, 4) cost effectiveness, and 5) broader impact. Analytical performance 'refers to the ability of a laboratory assay to conform to predefined technical specifications', while clinical performance 'refers to the ability of a laboratory assay to detect patients with a particular clinical condition or in a physiological state'. Clinical effectiveness 'refers to the ability of a test to improve health outcomes that are relevant to the individuals being tested, while cost effectiveness refers to the impact that the introduction of the test would have on an increase or reduction of use of resources'. Cost-effectiveness and clinical effectiveness are frequently addressed jointly. This joint approach to effectiveness is known as societal efficacy when costs are considered from 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 3 is mainly about the first two components: analytical and clinical performance. Evaluation of tests to detect infection focuses initially on analytical performance (3.1), which provides 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 (3.2) before the test can be recommended in a target population. *In Section 3, the critical statistical parameters for characterising diagnostic tests are set out. Studies to estimate these, which must be appropriately designed taking into account the framework set out in 2.4, will be considered further in Section 4.* 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.

3.1 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 (eg, molecule, antibody, antigen, or other; see Box 1 Section 2.1). 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 summarised by the Clinical and Laboratory Standards Institute (CLSI) (https://clsi.org) and the International Organization for Standardization (ISO) (www.iso.org). In this sub-section we briefly summarise key aspects of the assessment process to give context but refer 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).

3.1.1 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 summarised as a **coefficient of variation**, defined as CV = (Standard deviation/Mean) × 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.

3.1.2 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 Clinical and Laboratory Standards Institute (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 and Altman (Bland and Altman, 1995), is preferred (see Figure 3.1). 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.

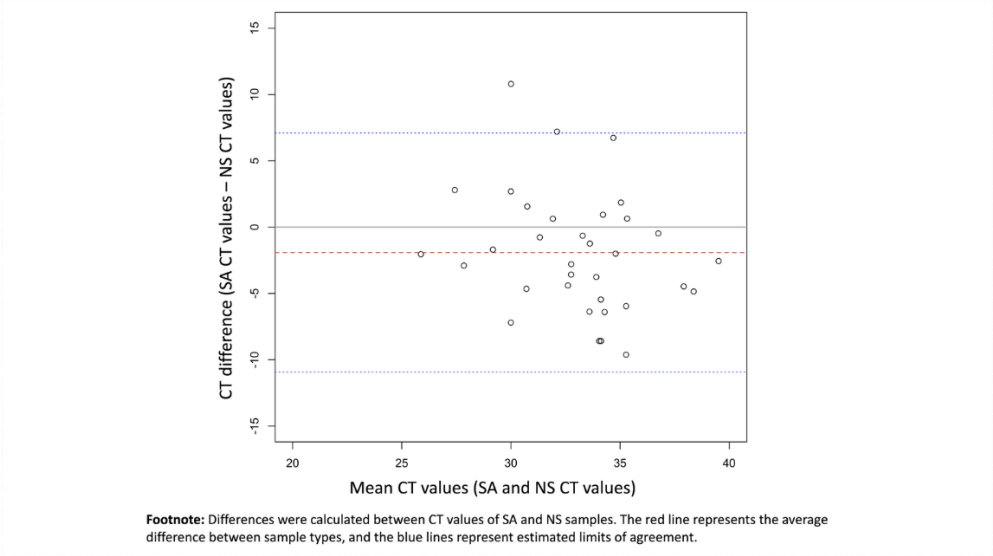


Figure 2 Bland-Altman plot of SARS-CoV-2-N1 CT values for SA and NS samples, Excerpt from Grijalva C *et al*, 2020, distributed under CC BY-NC-ND 4.0 license

3.1.3 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 around a specific analyte (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 *et al*, 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.

3.1.4 Analytical specificity

Some tests may pick up a range of pathogens, which are false positive findings if they are not the condition of interest**. Analytical specificity** assesses whether the assay is likely to give these false positive 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 false positive 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 (eg, rheumatoid factor, bilirubin, lipids), and to assess whether the ability to detect disease is compromised yielding false negative results.

3.1.5 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—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 95^{th} percentile value of a normal distribution with the calculated mean and standard deviation.

Many assays are not able to measure the smallest of 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 standard deviations 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.

3.2 Clinical performance

3.2.1 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 pre-selected groups of people already *known to have the target condition* and *known 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 are **field studies** that evaluate the performance of the test in a real-world setting, applied in 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 results of studies which adopted these two designs: those using pre-selected 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 false positive or false negative results) are often excluded from two-group studies, as only those with known disease state 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 4.2.2.

3.2.2 Diagnostic or clinical sensitivity and specificity

Diagnostic or clinical sensitivity and specificity describe the performance of the test 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 false positive rate (computed as 1-specificity) and false negative rate (computed as 1-sensitivity), either as ratios or absolute differences.

3.2.3. Predictive values

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 3.2.1). Presenting the impact of varying prevalence helps to determine potential utility of estimates of predictive values (PPV and NPV).

3.2.4 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 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 and Drug Administration, 2007).

3.2.5 Other aspects of test performance

Aspects of the tests other than their performance impact on their utility in a given setting. For example: **timing** (eg, how long it takes from obtaining biological sample to test result) and **human factors** that impact on usability and user errors need to be investigated (eg, 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.

Section 4: Study design for clinical performance studies

Section 4 delves into aspects of study design for clinical performance studies, the first of which is specifying the intended use or purpose of the new (index) test (4.1), which informs study design (4.2). Section 4.3 considers the implications of variations from the ideal study, and 4.4 addresses sample size. Study designs for wider purposes are briefly addressed in 4.5, and surveillance studies in 4.6.

4.1 Specifying the intended use or purpose

Specification of the intended use or use case is key to the appropriate evaluation to inform testing policy (Horvath *et al,* 2014; Doust *et al*, 2021). 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 recognised 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 *et al*, 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 2. The setting (eg, laboratory, hospital, general practice, surveillance-location) 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.

4.2 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.

4.2.1 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, gender 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 false positive compared to true positive. However, they may also impact on the stage and severity of the disease which will directly affect sensitivity.

4.2.2 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 (eg, hospitalisation) because they exceeded a decision threshold (eg, 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 an 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).

4.2.3 Timing of the tests

Whatever the test is actually measuring (eg, 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 to 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 that 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, too late and the levels might again have reduced to undetectable levels. This is the main reason that a diagnostic accuracy study for an 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.

4.2.4 Delivery of tests

Clear protocols should not only describe the target assay but also specify test delivery: **who** will obtain the specimens (eg, nurse, self-sample); **what** will be sampled (eg, blood, saliva); **how** will the specimen-be obtained (eg, finger prick, spitting into tube), stored and transported; and **when** (eg, within 7 days of first symptom). These elements should clearly relate to the proposed instructions of use for the test so that they reflect the way the tests will be used in practice. Within this protocol, information about how blinding of the results of the tests and the reference standard is carried out is also necessary to guarantee unbiased results (see Section 4.3.2).

4.2.5 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. The basic study design is still based around 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 randomised 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, randomising the order these specimens/tests are taken would likely prevent systematic bias (see Section 4.2.4).

4.3 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 **QUADAS-2** tool 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). See: https://www.bristol.ac.uk/media-library/sites/quadas/migrated/documents/quadas2.pdf (accessed 7 April, 2021).

4.3.1 Could the selection of patients have introduced bias?

The key consideration is whether the sample of patients in the study is representative of those in 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 3.2.1). 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 accuracy 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.

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

Interpretation/measurement of index and reference tests must be independent, ie, 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 re-interpretation of results and the retrospective evaluation of why there is 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 pre-determined during development of the test. Using the information from the study to define the threshold will lead to overestimation of the index test's performance.

4.3.3 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 false positives or negatives. 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 4.5).

As described in 4.3.2, 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 (Hagdu 1999)), substantial bias is likely, typically artificially increasing the estimated accuracy of the index test.

4.3.4 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 (eg, 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, randomising the order of within-person tests will be necessary. When the specimens required are substantially different, eg, nasopharyngeal swab versus saliva, then it is feasible to take these different samples at roughly similar times without the order in which they are taken affecting the results.

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 (eg, if this requires a certain duration of follow-up). In these scenarios it is important to highlight which reference standards were used, on which participants and mention potential advantages/disadvantages of using each reference standard.

As discussed in Section 4.2.1, patient selection is critical as 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.

4.4 Sample size issues and incorporating or evaluating uncertainty

The choice of methods and approach to sample size estimation for studies assessing clinical performance will vary depending on the objective/s (Obuchowski, 1998). For example, if the regulatory focus specifies a required performance (eg, 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 could 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 randomised trials is normally carried out based on expected difference in parameters (eg, sensitivity or specificity), probability of type one error and power (alpha and 1-beta respectively). Korevaar (Korevaar *et al,* 2019) describe a clear framework for such studies and 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 *et al*, 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.

4.5 Study designs for other relevant clinical studies of diagnostic tests

So far we have focused on relevant study designs required to determine clinical accuracy. There are several other clinical questions critical for our understanding of test performance which fall outside the clinical accuracy setup.

4.5.1 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 understand the natural history of the disease post-infection and to determine 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 (eg, 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.

4.5.2 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 the benefits (often counted as cases detected) and harms (for example, unnecessary isolation, disinhibition from false negative test-results leading to potential for increased transmission, lost income). New tests, particularly point of care tests, may change radically access to testing, leading to differences in 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 *et al*, 2012) via intended and unintended effects—and can help in planning evaluations which need to be undertaken. Effects can be categorised under four main headings: (a) direct test effects on the patient (eg, 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 (eg, reducing time to diagnosis and treatment); (d) influencing patient and clinician perceptions and behaviours (eg, willingness to undergo procedures, the impact of test results on patient behaviour, defensive medicine).

Large cluster randomised trials comparing different testing options can be difficult logistically or in policy-terms. Randomised 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 testing can have.

4.6 Surveillance studies

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

4.6.1 Confidentiality in surveillance studies

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 the following:

(1) Unlinked anonymous testing (UAT) makes use of a residue or aliquot from a testable biological specimen given for other reasons. UAT 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. UAT is designed so that there can be no deductive disclosure about individuals: hence, only minimal information about the person whose specimen is tested (such as gender, broad age-group, and region) is retained with the surveillance-specimen.

(2) Consented attributable linkage of 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 (eg, nasopharyngeal swab in SARS-CoV-2).

(3) Consent for non-attributable 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 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*, 1999; Gore *et al*, 1995; Yirrell *et al*, 1997; Hutchinson *et al*, 2000; White *et al*, 2015). High volunteer rate matters, which non-attribution 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 questionnaire about risk-behaviours affords privacy and engenders frankness.

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, are being obliged to disclose to NHS Test & 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, phone number and email address are also reported to NHS Test and Trace.

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

Section 5: Lessons learned from evaluating tests for COVID-19

Sections 1 to 4 are intended to apply to a wide range of diagnostic tests in a variety of diseases and situations, but the main motivation for this report has been the challenges posed by COVID-19 pandemic. After an introduction to testing for SARS-CoV-2 antigen or antibodies, Section 5 follows the order of Section 4, with specific detail on intended use cases, study designs, participants, tests, target conditions and reference standards, concluding with general recommendations on test evaluation methods.

5.1 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 2.1). 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 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 (Cevik *et al*, 2021). Initial immune responses (see Section 1.2) are of IgA and IgM antibodies, with IgG appearing later and lasting longer (Post *et al*, 2020). 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.

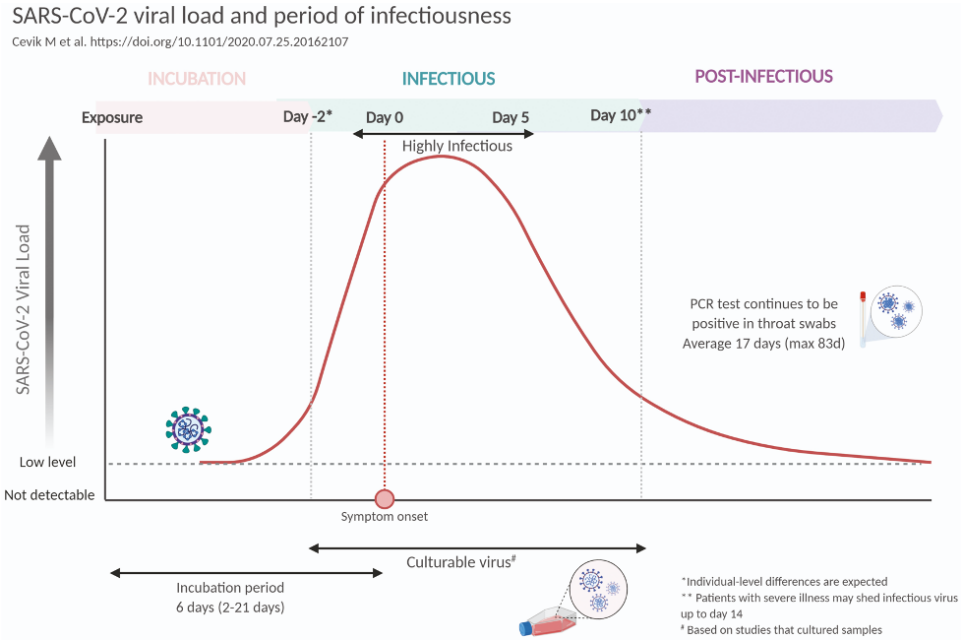


Figure 5.1A. Schematic of viral levels during a COVID-19 infection (source Cevik *et al*., 2021 –Copyright Oxford University Press, reproduced with permission)

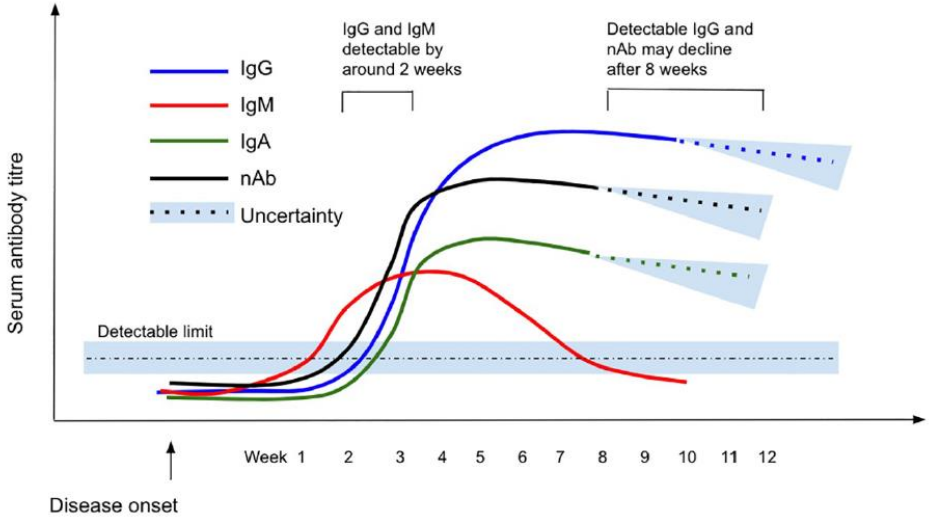


Fig 3. Schematic showing the scale of IgG/IgM/IgA/Neutralising Ab response over time from disease onset. Note that the y-axis is illustrative only and therefore no scale is given: this figure gives an indicative overview of findings from all included studies with relative peaks and decline indicated.

https://doi.org/10.1371/journal.pone.0244126.g003

Figure 5.1B. Schematic of antibody levels during a COVID-19 infection (excerpt from Post *et al*, 2020, distributed under CC BY license)

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*, 2020), opportunistic epidemiological studies of communities (such as outbreaks on cruise ships (Hung *et al*, 2020; Mizumoto *et al*, 2020)) and through insights gathered via cross-sectional studies using diagnostic tests (Deeks, Dinnes *et al*, 2020). One initial challenge has been to foresee and undertake the research required to acquire this understanding, when no single party has central oversight and control, and to update understanding as new information emerges.

5.2 Intended use cases

Key to the appropriate evaluation of tests to inform testing policy has been the specification of "intended use cases" (see Section 4.1). In COVID-19 testing, it is 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 be used to determine patient management, control disease, and/or for surveillance purposes.

Several organisations (eg, Foundation for Innovative New Diagnostics, 2021) have 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 5.1) emphasising the importance of undertaking evaluations of tests for each intended use; and monitoring their performance during implementation.

During the pandemic, new potential intended use cases have emerged as new considerations have been made about how best to tackle disease spread, whereas others—originally thought to be important—have been found to be unnecessary, impractical or impossible (Example 5.2).

Example 5.1 – 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 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 percentage 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*, 2020).

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 were consented 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 5.8.

Example 5.2 – 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 S, 2020; Royal Society of Medicine, 2020).

Initial point-of-care antibody test performance was deemed inadequate (Adams ER *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.

5.2.1 Intended use cases for molecular and antigen tests

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

a) symptomatic people to diagnose COVID-19.

However, they also have 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.

The target condition is thus either SARS-CoV-2 infection (if asymptomatic) or COVID-19 (if symptomatic). In addition, mass testing has been proposed to identify individuals who are infectious rather than just infected. Problems in considering infectiousness as a target condition are discussed in Section 5.6.2. 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 are several different types of test which vary 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 specimen and result. Most tests require a throat and/or nasal swab, some are being trialled on saliva samples. Decisions about use of different tests thus depends on performance, but also on cost, speed, capacity and accessibility.

5.2.2 Intended use cases for antibody tests

Antibody tests identify whether an individual has developed antibodies to SARS-CoV-2. These tests have been 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 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.

5.3 Study designs

Section 3.1 outlines the limitations of analytical studies which assess the properties of a test on samples in laboratory settings, and Section 3.2.1 describes 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.

5.3.1 Regulatory requirements for evidence have varied

New tests have been developed at pace to address the emergency need for diagnostics during the SARS-CoV-2 pandemic. The standard process described in Section 3.1 of undertaking laboratory studies of the key analytical properties of new tests prior to assessing their accuracy in real world field settings has been followed. However, evidence from analytical studies has been 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 they begin to be used for their intended uses have often followed later, sometimes alongside their implementation, but sometimes not at all.

Evidence requirements are not consistent across regulators; and have changed during the pandemic (Shuren *et al,* 2020). The CE-IVD marking used across the UK and 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 (EC) 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 FDA has a more rigorous process but has frequently allowed tests to market based on Emergency Use Approvals reliant on similar analytical performance evidence (see Example 5.3). For a period during 2020, all antibody tests were allowed to be marketed in the USA without restriction, a decision which allowed poor performing tests onto the market (Shuren *et al,* 2021).

Example 5.3 – 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).

5.3.2 Analytical and two-group study estimates of sensitivity and specificity

Manufacturers usually report estimates of test sensitivity and specificity in their Instructions for Use (IFU) documents and separate out 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 which 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 emphasised 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 5.4), 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 pre-selected specimen banks of known disease status or collected prospectively for particular use cases (see Example 5.4).

Example 5.4 – 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 (A), 2020). 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*, 2020). 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 3.2.1). See also Example 5.9.

Later non-selective 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).

5.4 Participants

Studies of COVID-19 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; the specificity in 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 false negatives in antigen tests which cannot detect lower levels of the virus; testing before antibodies rise will increase false negatives for antibody tests; and testing when the prevalence of infection is low will increase the proportion of test-positives that are false-positives.

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 versus asymptomatic groups (see Example 5.1) (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 which have no or low antibody responses; and will have lower accuracy when used soon after symptom onset compared to later (Example 5.5). If antibody response relates to disease severity, the performance of the test will differ between those with 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 false positive results, if the test fails to distinguish between similar viruses. In

the first Cochrane review of antibody tests, the false positive rate for IgG tests in six studies (n=396) that recruited individuals suspected of having COVID-19 was 2.0% (95% CI: 0.4 to 9.0) compared to 0.8% (0.2 to 2.4) in 10 studies (n=2614) that included currently healthy participants, and 0.8% (0.3 to 2.2) in 10 studies (n=2633) on pre-pandemic cohorts (Deeks, Dinnes *et al*, 2020). Notice, however, that the confidence intervals here are too wide for inferences to be drawn.

5.5 Index test

5.5.1 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 5.2). When tests were used at a later time point, higher levels of sensitivity were obtained (see Example 5.5). Some antibody tests may have been inappropriately abandoned due to their lack of accuracy in the early time period.

As the pandemic progresses, waning of antibody responses will affect the ability of surveillance studies based on antibody tests alone to identify previously infected cases, as they will only identify those in whom antibodies are still detectable. Evidence is beginning to accumulate on the likely duration and variability in antibody response between tests and between individuals (Ward *et al,* 2020). Equally, it may become difficult to identify those with responses to infection from those with responses to vaccination (Ward *et al,* 2021).

5.5.2 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 RT-LAMP test which has been compared on real and spiked samples (Example 5.6), and on swab and saliva samples, with and without RNA extraction stages, was found to vary between 70% and 95% (Example 5.7) (Department of Health and Social Care (A), 2020).

There are also many studies which have used specimen types outwith the manufacturer's approval (such as saliva or viral transport media when dry swabs are required), or blood samples taken in different ways (venous versus skin prick for antibody tests – see Example 5.8). 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.

Variation in Innova test's reported sensitivity illustrates the potential for combinations of population, the tested material, and testing operative to impact on sensitivity of an antigen test (Example 5.9).

Example 5.5 – 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 Emergency 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 [IQR] of 7 [4, 11] days after their first test.

The Cochrane review (Deeks, Dinnes *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% (22 to 39) 23 studies 568 samples	67% (58 to 74) 22 studies 1200 samples	88% (84 to 92) 22 studies 1110 samples	80% (72 to 86) 12 studies 502 samples	87% (80 to 92) 4 studies 252 samples
IgG/IgM	30% (21 to 41) 9 studies 259 samples	72% (64 to 80) 9 studies 608 samples	91% (87 to 94) 9 studies 692 studies	96% (91 to 98) 5 studies 52 samples	78% (66 to 86) 2 studies 53 samples

Example 5.6 – Differences between real and spiked samples

The DHSC evaluated the RT-LAMP test (see also Example 5.7) on saliva and swab samples (Department of Health and Social Care (A), 2020). 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 versus spiked samples, when stratified by viral level (as defined by the Ct value from the accompanying PCR-RT 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 5.7 – 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 and Social Care (A), 2020). (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	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%)*

** this is calculated excluding the 59 spiked samples mentioned in Example 5.6.*

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 percentage points (95% CI: 18% to 32%). False negatives also increased by 15 percentage points (95% CI: 7% to 23%) when using RNA from saliva samples instead of RNA extracted from swabs.

5.5.3 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 utilise 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 5.8).

Example 5.8 – 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 are 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 and Healthcare products Regulatory Agency, 2020).

The performance of point-of-care tests 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*, 2020).

5.5.4 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 requires evaluation: but currently, there are few empirical evaluations of the performance of serial antigen testing strategies. Many estimates of predicted performances are based on naïve extrapolation from a single test use, under an independence assumption.

Naïve **Bayes** estimation of serial test performance assumes independence between successive tests. Assuming independence allows multiplication of the probabilities of false negatives (or false positives) as a means for estimating serial performance. For example, if independence held and infection-status does not alter, applying test X with a false negative rate of 30% twice would yield an overall 9% false negative rate (0.3 x 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, Raffle *et al*, 2021). However, as the false negative-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 5.6.3 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*, 2020, Quilty *et al*, 2021). It is important that robust randomised empirical evaluations of serial-testing strategies are undertaken as any modelling of the underlying biology will necessarily rely on simplifying assumptions.

Example 5.9 – Differences in estimates of sensitivity: Innova test

There have been 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 have shown 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) have lower sensitivity than those in symptomatic groups. There have also been differences between whether fresh swab samples are tested versus testing done on frozen samples or by use of the viral transport media; and whether the tests have been run/read by laboratory professionals, healthcare workers 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*	Not stated	72/ 75	96% (89 to 99)
PHE [2]	SARS-CoV-2 positive patients	Hospitalised 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*	HCW	156/223	70% (63 to 76)
PHE Phase 4 [2]	Symptomatic	Test-and-trace centre	Dry swab*	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*	non-HCW	28/70	40% (28 to 52)
Uni B'ham [5]	*Asymptomatic*	*Student testing*	*Dry swab**	*non-HCW*	*2/62†*	*3%† (1 to 16)*

** Tested according to manufacturer's instructions*

†This study sampled 10% of non-cases. Total COVID-19 numbers, sensitivity and its confidence interval are computed reweighting for the study design (see Example 5.11)

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 (A) 2020; [2] Peto et al 2021; [3] Dinnes et al 2021; [4] University of Liverpool 2020; [5] Ferguson et al 2021.

5.6 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.

5.6.1 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 test-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 5.2.1) there are three different target conditions: (a) considers whether individuals are currently infected; (d) whether they have previously been infected; whereas for (b), (c) and (e) the target condition is the presence of antibodies. See Example 5.10.

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 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 pre-pandemic 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 5.10 – 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 and Social Care (B), 2020). 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 B, 2020). The manufacturer assessed the test in a study of known antibody positive and known antibody negative samples (see Example 5.4). 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* 2020). 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.

5.6.2 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 have been identified as more infectious than others (ie, "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 are 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. Direct evidence of transmission to secondary cases clearly indicates infectiousness, but absence of transmission does not indicate non-infectiousness, particularly when people have been isolating (and thus preventing transmission which otherwise would have occurred).

It has also been argued that infectious people must have viable virus (in that it can replicate). Viral viability has thus been 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, eg, Bullard *et al* 2020 found 26/90 (29%; 95% CI 20 to 39) whereas Singanayagam *et al* 2020, found 134/324 (41%; 95% CI 36 to 47))

Cycle-threshold 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. There are, however, very few

studies which have attempted viral culture in individuals receiving rapid antigen tests, (eg, Schuit *et al* 2021). Put simply, maximum Ct values (ie, minimum viral loads) identified in studies above which secondary cases or viral culture do not occur have been 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 *et al* 2007; the second is that "infectiousness threshold" needs external validation before being adopted more widely; and third that the empirical data display a continuity of decreasing risk with increasing cycle thresholds without any clear lower bound (Singanayagam *et al* 2020 – also see Example 5.12). However, policy appears to have been constructed from stating binary thresholds. For example, 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 states "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 (B), 2021).

The regulator has 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 and Healthcare products Regulatory Agency (B), 2021). This is more inclusive than infectiousness, but equally there is no clear reference standard at present that can clearly differentiate between current and recent or previous infection.

5.6.3 Establishing the performance of RT-PCR

The RT-PCR test has been the primary diagnostic test for SARS-CoV-2 infection used globally from the start of the pandemic and has become established as the reference standard against which other tests are compared. However, questions have been raised concerning the performance of the RT-PCR test, particularly its false positive 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 false positive rate with RT-PCR by considering the total positive rate when disease prevalence is low (see Example 5.11).

Example 5.11 – Estimating the performance of RT-PCR without a reference standard

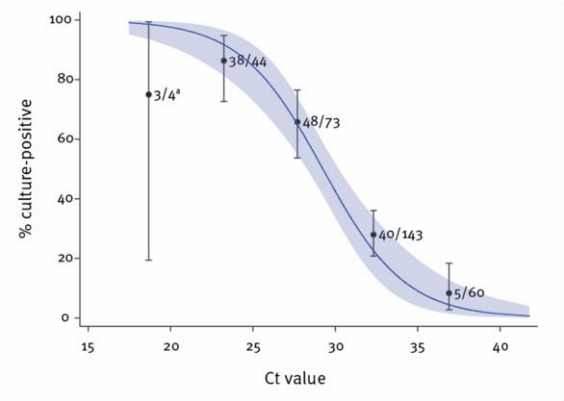
UK population prevalence surveys using RT-PCR have shown test positivity rates from the Office of National Statistics Community Infection Survey in England of 0.44% (95% credible interval: 0.22 to 0.76) in August 2020 (Office for National Statistics, 2020), and 0.077% (95% confidence interval: 0.065, 0.092) from the 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 negative predictive value of RT-PCR has been 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 false negatives. A review of such studies showed variation between 2% and 54% of negative PCR tests being false negatives, 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 (the cycle threshold or 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 has become a key issue for antigen tests but is hampered by an absence of standardization of Ct values between machines. Studies to understand the relationship between Ct values and viral load are 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 5.12). The relationships have then been estimated using linear regression, some regressing $\log_{10}$ viral load on Ct value (eg, Mitja *et al*, 2021) others Ct value on $\log_{10}$ viral load. Rarely is 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 has been paid to use of standard measurement comparison methods (such as Bland-Altman methods, see Section 3.1.2) to estimate systematic differences in Ct values between different machines. Nor are there many studies of measurement error in Ct values.

Example 5.12 – Relationship between viral load and viral culture

Public Health England 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 A *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 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). 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, have shown 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.



(Bars represent 95% confidence intervals)

Excerpt from Singanayagam A et al., 2020 distributed under CC-BY 4.0 license

5.7 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**). Public health discussion often naturally focuses on the sensitivity as it describes the proportion of cases that will be detected, and the

specificity as it describes the risk of false positives, 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 (false negatives and false positives) are explained clearly to avoid confusion. Good practice is to ensure that the probabilities related to false positives and false negatives are explicitly explained, rather than relying on the public's understanding of sometimes arcane statistical terminology (see Example 5.13). For example, the phrase "false positive rate" is sometimes used for both for 1-specificity and 1-PPV, which can cause confusion. When infection events are rare (ie, when prevalence is low), these two probabilities differ considerably and, if confused, give seriously wrong impression. When the prior probability of infection is very low, even tests which have exceptionally high specificity can give more false positives than true positives (see Example 5.14).

Example 5.13 – Poor communication about test performance sent to schools

The Department for Education guide to schools (NHS Test and 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, Gill *et al*, 2021) 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 (positive predictive value). 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, **negative predictive values** 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 disease is 1 in 1000, those testing negative on Innova have a post-test probability of 99.94%, which to many appears to indicate very low risk of disease. However, this must be compared with the chances of disease 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** 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 5.14 – 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 false-positives may be high, as the RSS COVID-19 Taskforce had forewarned (Royal Statistical Society COVID-19 Taskforce, 2021).

The proportion of LFT-positives that are RT-PCR negative depends on the prevalence. Using the figures from Liverpool (Example 5.8) for the performance of Innova test shows that the positive predictive value ranges from 80% if 1 in 100 are infected, to 29% if 1 in 1000 are infected and 4% if only 1 in 10,000 are 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 RSS COVID-19 Taskforce anticipated that half would be asymptomatic infections.

Scenario A - 1 in 100 pupils have Covid-19 infection

	Covid	No Covid	Total	**Percentage with 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

	Covid	No Covid	Total	**Percentage with infection**	
LFT+	400	999	1,399	LFT+	28.6%
LFT-	600	998,001	998,601	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

	Covid	No Covid	*Total*	**Percentage with infection**	
LFT+	40	1,000	1040	LFT+	3.8%
LFT-	60	998,900	998,960	LFT-	0.006%
Total	100	999,900	1,000,000	Overall	0.010%
prevalence	***0.01%***				

Data for the first two weeks of testing were made available at the end of April 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 and Social Care (C), 2021). As predicted, the proportion of positive LFTs that were false was high, with 605 negative and 428 positive (17 tests were void): a positive predictive value 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 three weeks after the mass testing in schools commenced.

5.8 Advanced study design issues

5.8.1 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 randomised groups (Takwoingi *et al*, 2013).

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 has 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 are needed: there have been 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 have been done using alternative samples to swabs, such as viral transport media (eg, Pickering *et al,* 2021) which may be outside the specimen-types for the test's approved use.

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, D). Then each patient-volunteer can be randomised 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 randomisation to one of six possible pair-wise comparisons: AB or AC or AD or BC or BD or CD. Good practice would also involve randomising swab-order as it may matter.

5.8.2 Using efficient designs

The prevalence of SARS-CoV-2 infection has generally been less than 0.5% but with 4-fold increase or decrease also observed during pandemic waves. Hence, prospective studies have 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 5.15). Whilst positive and negative 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 5.15 – Use of sampling for efficient designs

One study has used sampling to provide efficient estimates of the performance of the Innova lateral flow assays in the UK (see Example 5.8). 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 7187 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 (TTP) Service). 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.

5.9 RECOMMENDATIONS on 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 **multisite 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 centralised processes for ethical and study-protocol approvals.

Section 6: Regulation

Sections 1-4 outlined the scientific basis for evaluation of diagnostic tests, with Section 5 having illustrated these in the context of SARS-CoV-2. Section 6 looks at the regulatory implications, with a particular focus on the UK: 6.1 sets out the current and evolving regulatory position; 6.2 sketches new regulatory approaches that are emerging.

6.1 Regulatory context

The Medicines and Healthcare products Regulatory Agency (MHRA), an agency of the UK Government, regulates medicines, medical devices and blood components for transfusion in the UK. There are differences in the law that applies for oversight of medicines versus oversight of devices, including how regulation is funded. Medicines regulation is funded by licence fees, paid by the manufacturers, whilst the regulation of devices is funded by the Department of Health and Social Care.

In advance of the COVID-19 pandemic, revisions were underway to the regulatory framework that applies to IVDs. European Union law was revised from a Directive (European Parliament and Council, 1998) to a Regulation (European Parliament and Council, 2017), and UK law is also in the process of revision post-Brexit.

Under current procedures, the commonly applied approach to certification is to affix a CE mark, with the manufacturer (not regulator) primarily responsible for this. The responsibility of the manufacturer, and responsibilities of the bodies they notify, are outlined in the 1998 Directive.

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 implications of the current approval process are that IVDs are not being independently scrutinised, as the process is one of notification – not assessment. Notification allows for inconsistency in the evidence available to support the use of different tests and means that the CE-IVD mark cannot be taken as evidence of independent review that a test is fit for use in a clinical setting.

Currently tests which are to be administered by members of the public and not healthcare professionals do require evaluation by a notified body to obtain a CE marking that allows self-use. During the COVID-19 pandemic, regulators have provided time limited “exceptional use authorisations” for IVDs—this is a route to

approval separate from the CE mark. Devices authorised by this route have been publicly listed (Medicines and Healthcare products Regulatory Agency, 2021), whereas those approved by usual routes have not been.

Several authorities including the MHRA provided for **exceptional use authorisation** of IVD devices in the COVID-19 pandemic. In the UK, exceptional use has authorised particular IVD devices on a time-limited basis for use by the public for self-testing, bypassing the more strenuous approval process usually required. Devices granted an exceptional use authorisation can be sold to the NHS and for use in social care.

Lists of devices authorised and no longer authorised by this route in the UK and of companies manufacturing them are provided here: https://www.gov.uk/government/publications/medical-devices-given-exceptional-use-authorisations-during-the-COVID-19-pandemic (accessed 13th May 2021).

Further guidance on exceptional use authorisation is provided here: https://www.gov.uk/guidance/exemptions-from-devices-regulations-during-the-coronavirus-COVID-19-outbreak (accessed 13th May 2021).

Except for such exceptional use authorisations, a confidentiality clause has ruled out a more comprehensive public register of devices and has also been cited against responding to Freedom of Information Requests in the UK.

The confidentiality clause is revised by the new EU regulatory framework. In April 2017, the EU had brought into force a new Regulation on *In Vitro* Diagnostic Medical Devices, which stated: "a fundamental revision [to the law] is needed to establish a robust, transparent, predictable and sustainable regulatory framework for in vitro diagnostic medical devices which ensures a high level of safety and health whilst supporting innovation."

Post-Brexit, the UK also plans to enact new UK laws. Under the MHRA's corporate plan (2018-2023) and as allowed for in the Medicines and Medical Devices Bill, the UK will implement *In Vitro* Diagnostic Medical Regulations by May 2022, and refer to the preceding EU Directive, 98/79/EC, in the interim. The EU similarly has an extended transitional period in its own Regulation, meaning that manufacturers in member states have until the 26th of May 2022 to update their technical documentation and processes, and may (as is the case in the UK) still refer to the preceding EU Directive, 98/79/EC (1998) at the present time.

6.2 New approaches

6.2.1 Target Product Profiles (TPP) for COVID-19 tests.

The purpose of a TPP is to "outline the desired 'profile' or characteristics of a target product that is aimed at a particular disease or diseases." TPPs "state intended use, target populations and other desired attributes of products, including safety and performance-related characteristics."

A TPP "provides a common foundation for the development of tests and contains sufficient detail to allow device developers and key stakeholders to understand the characteristics a test must have to be successful for the particular intended use. Included is a description of (1) the preferred and (2) the minimally acceptable profiles based on the intended use, setting of use, and intended user, with respect to the performance and operational characteristics expected of the target products."

It can be challenging to ascertain the performance for a test that will make it fit for intended use, and research methods to support this task are required.

A number of authorities, including the UK's MHRA, have published Target Product Profiles (TPPs) for COVID-19 tests, and these are an important tool.

The role of the MHRA in defining suitable reference standards is an issue for consideration. Current TPPs define the properties required for a reference standard to be used to establish test performance, but they do not state reference standards are deemed to meet these criteria. Greater standardisation of research could be achieved by consensus on suitable reference standards which could be written into future TPPs.

6.2.2 Changes in the law applying to IVDs

Regulation (EU) 2017/746 on *in vitro* diagnostic medical devices will be fully implemented in EU countries by 23 May 2022. The Regulation includes new compulsory arrangements for transparency and limits on self-certification.

Implications vary according to class of device, ranging from self-certification by the manufacturer for the lowest risk. But for many different IVDs for infectious diseases, manufacturers will be required to summarise the main safety and performance aspects of the device and the outcome of the performance evaluation in a document that should be publicly available; and an appropriate level of involvement of a notified body (that is: the regulator) will be compulsory.

Further, the above EU Regulation notes that harm to the patient, their offspring, or disease management, may be caused by misdiagnosis in the use of IVDs. The Regulation assigns duties to the 'notified body' for independent verification and testing, for audit of the quality management system, and assessment of the manufacturer's technical documentation on the basis of further representative samples.

Conformity with European law will be essential for sales of IVD devices into the EU market in future. The MHRA therefore plans to harmonise the UK's domestic law with EU regulation.

These changes offer a real opportunity for major strengthening of the regulation of diagnostic medical devices to bring them on the level of the evidential requirements applied to medicines. Additionally, there are areas where the law may be strengthened: for example, so that the public have access under law to the manufacturer's data that underlie IVD registration.

6.3 RECOMMENDATIONS on regulation matters

1. **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.
2. 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**.
3. 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.
4. 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 liberalised behaviour to deprivation of liberty, should be considered.
5. Evaluation of the **impact of tests** should ensure that both **intended** and **unintended consequences** are considered. Some consequences will not be evaluable before test implementation, so that **post-marketing surveillance** for a new intended use requires ongoing assessment.
6. 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.

Section 7: Information to be in the public domain

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 policy makers 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-organised, transparent reporting of all stages of research has been established for randomised trials of interventions, encompassing:

- prospective registration of studies on public registers to ensure all research studies can be identified regardless of their findings (to prevent publication bias) (Simes, 1986).
- prospective publication of study protocols and statistical analysis plans to provide evidence of the original design of the study and specification of the outcomes, patient groups and analyses (to distinguish pre-specified analyses from potential data-driven analyses) (Chan *et al*, 2018).
- peer review of protocols and study reports (to validate the science and enhance the quality of reports) (Yordanov *et al*, 2015).
- timely publication of full study methods and study findings, encouraging open access publication (to provide full information to the public and clinicians on effectiveness) (Dwan *et al* 2013).
- provision of access to data sets (to enable study findings to be confirmed and data presentation to be harmonized) (Taichman *et al*, 2017).

The same emphasis has not yet been applied to test evaluation studies but, given that similar concerns arise (particularly selective publication and data driven analyses), these principles equally apply in IVD research and should be endorsed.

Different stakeholders have different information needs: 7.1 outlines what the public, patients and clinicians need to know, 7.2 considers the perspective of policy makers and 7.3 addresses researchers and study participants.

7.1 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 make an informed choice, and that they appreciate potential downsides as well as benefits of testing. In the past, 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 jeopardised" (Science and 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 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 false positive and false negative 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 understand is thus essential (see Examples 5.13 and 5.14).

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** (eg, 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 COVID-19 lateral flow antigen tests initially used in the UK, there have been 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 5.7, negative test results do not rule out infection or infectiousness. It is essential that the public are 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.

7.2 What policy makers need to know

In pandemics, decisions may have to be hedged about which tests to purchase because the available tests have been evaluated in more limited contexts-of-use than eventually apply. As with vaccines, it may be prudent for governments diversify their portfolio of test purchases and for changes to be made in the light of new evidence.

The approach which has been adopted and endorsed by the UK Government for review of screening programmes contains processes and information of relevance to more general use of tests, particular for consideration of 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, 2014) 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 false negative result) thereby missing the correct diagnosis; or giving a positive result when the result should have been negative (a false positive 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 false positives and false negatives 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 false positive. As infection levels drop, fewer cases will be detected, whereas the likely harms through false positives 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 5.14). 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 on policy decisions.

Tests alone do not improve patient and population outcomes: it is the interventions which follow that create benefit. Implementation of testing therefore needs to be linked to 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 with SARS-CoV-2 infection can and do isolate is essential for preventing onward transmission of infection.

Policy makers thus require access to all study findings, which can usefully be summarised in **systematic reviews** that identify, appraise and synthesise all relevant evidence; and assess the strength of the evidence. 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. Availability of pre-prints from on-line archiving services has revolutionised the ability to achieve this.

Separate evidence reviews are required for each intended use – each with assessment of the strength of evidence, taking note of the findings and the inherent uncertainty in the estimates (results must be presented with confidence intervals or equivalent) and 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 of 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 PMM *et al*, 2015) and systematic reviews (PRISMA-DTA) (McInnes *et al*, 2018; Salameh J-P *et al*, 2020) should be followed to ensure that all relevant details are included.

7.3 What researchers and study participants need to know

Research studies may not all be made public, or key details omitted, either for commercial reasons or academic interests. For example, due to the commercial confidentiality clauses in the contracts with the manufacturers, HMG-sponsored evaluations of point of care antibody tests did not name all the test kits evaluated: indeed, none of the nine in the antibody study (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 and 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 maximise the accuracy of the reference standard or assess sequelae.

7.4 RECOMMENDATIONS on transparency matter

1) **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**.

2) **Expert peer review** of study protocols and final reports by **subject-matter** (eg, clinical, public health, laboratory) and **methodology experts** is recommended.

3) **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.

4) **Post hoc analyses** should be limited; and clearly identified as **exploratory**.

5) **Study reports** and results should be made available **publicly** in a **timely manner**.

6) Field and clinical evaluation studies require **ethical approval** and **fully informed consent** as outlined in the **Good Clinical Practice Guidelines**.

Appendix 1: Examples of major infectious diseases

A1.1 Pandemic and seasonal influenzas

Prior to swine-flu in 2009 (influenza A/H1N1), which was a mild disease, three pandemic influenzas occurred in the 20th century: Spanish flu in 1918 (A/H1N1), in three waves, first and worst because it arose before penicillin or modern virology; Asian influenza in 1957 (A/H2N2); and Hong Kong influenza in 1968 (A/H3N2), see Greenwood, 1918; Kilbourne, 2006; Birrell *et al*, 2016.

After swine-flu, pregnant women were again recognised as a vulnerable group and advised to be immunized against seasonal influenza in the UK. Antibodies to previous years' seasonal influenza (A or B), by either exposure or immunization, can afford partial protection to older individuals. Such protection is less likely to apply in pandemic influenza because the virus is new, having arisen from genetic re-assortment with animal influenza A viruses, so that even cross-reactivity may be minimal. However, cross-reactivity did help senior citizens during swine-flu (Miller *et al*, 2010, Bird, 2010).

A1.2 Hepatitis viruses

There is an alphabet of hepatitis viruses (A, B, C, …) which differ in how they are transmitted, the risk they pose for developing chronic hepatitis, and in their potential for protection by immunization.

Hepatitis A virus: The main route of transmission for Hepatitis A virus (HAV) is faecal-oral through contaminated food or water. There is no carrier state so that the UK's blood supply is not screened for HAV. However, as viraemia develops before HAV's symptom onset, transfusion transmission of HAV has been reported sporadically, see da Silva *et al* (2016) and http://www.transfusionguidelines.org/transfusion-handbook/5-adverse-effects-of-transfusion/5-3-infectious-hazards-of-transfusion. Immunization protects us against HAV.

Hepatitis B virus: Identified in 1967, Hepatitis B virus (HBV) is highly infectious. About 10% of those infected remain infectious HBV carriers and are at risk from late liver sequelae. Besides being blood-borne, HBV is transmitted sexually, from mother-to-child, in saliva and even sweat. Immunization (including in infancy) now protects against HBV and is recommended for those at increased HBV-risk such as healthcare workers (mandatory in UK) or injection drug users; and in the developing world (Viviani *et al*, 2008; Van Damme, 2016). From the early 1970s, the UK's blood was protected against HBV.

Hepatitis C virus: Transfusion-transmitted infections persisted after HBV was identified: so-called non-A, non-B hepatitis. The Nobel Prize in Physiology or Medicine was awarded in 2020 to Alter, Houghton and Rice for seminal experiments in 1987 and 1988 which led to the discovery of the Hepatitis C Virus (HCV) as the causative agent of non-A, non-B hepatitis. From 1991, the UK's blood supply has been protected against HCV by progressively improved screening (Spearman *et al*, 2019).

Sexual transmission of HCV does occur, but infrequently. Unmitigated, the risk of vertical transmission from HCV-carrier mother-to-child is around 7%. Unlike HBV, only a quarter of those who are HCV-infected clear the virus spontaneously (Hutchinson *et al*, 2005). In Europe and USA, injection drug use is the major risk-factor for HCV infection. The USA's Centers for Disease Control and Prevention recommend HCV screening at least once in their lifetime for all adults, unless HCV-prevalence is below 0.1%, see https://www.cdc.gov/mmwr/volumes/69/rr/rr6902a1.htm. Vaccine against HCV is an unlikely prospect.

The 21st century has seen remarkable pharmaceutical development of directly acting antiviral (DAA) treatments which clear HBV and HCV in a high proportion of infectious carriers (Spearman *et al*, 2019; Hutchinson *et al*, 2020). Until the advent of DAAs, HCV-genotype (eg, 1 versus 4) determined the duration of earlier, less successful interferon-based therapies (Struble *et al*, 2019).

A1.3 Human immunodeficiency virus (HIV):

Until the virus which caused Acquired Immunodeficiency Syndrome (AIDS) was isolated in 1983, protection of the blood supply against HIV/AIDS was initially by deferral of at-risk donors (eg, men who have sex with men, injection drug users, persons from high-AIDS-prevalence countries). Transmission-routes for HIV are blood-borne, sexual and mother-to-child (including via breast milk). Spontaneous HIV-clearance is almost unknown.

By 1984, first and second-generation HIV antibody tests had been developed; with antigen testing soon to follow (Alexander, 2016). The UK's blood supply was protected from October 1985.

Antigen testing gives better protection of the blood supply than antibody testing because there is inevitably a delay between being infected (antigen presence) and immunological response (antibody formation). In the window-period between HIV-infection and antibody formation, individuals will test negative for HIV antibodies despite being HIV-infected (antigen positive). For this reason, protection of the UK's blood supply relies upon antigen testing for both HIV and HCV. Moreover, in the light of HIV, stronger directives apply to the UK's regulation of tests for blood-borne infectious diseases because the blood supply needs to be protected. Not until 2014 were home HIV-test kits approved for use in the UK (Public Health England, 2014), earlier attempts having been barred (Gore, 1992).

Currently, there is no authorised vaccine against HIV. However, since 1996, highly active antiretroviral therapies (HAART) have transformed HIV survivorship (Collaborative Group on AIDS incubation and HIV survival including the CASCADE EU Concerted Action 2000; Babiker *et al*, 2002); minimized mother-to-child HIV-transmission; and can be used for pre-exposure prophylaxis. Affordable Polymerase Chain Reaction (PCR)-testing allows treatment decisions to be informed by prior knowledge about the patient's resistance to specific HAARTs (Detels *et al*, 1998; Babiker *et al*, 2002).

A1.4 Variant Creutzfeldt-Jakob Disease (vCJD):

Animal experiments demonstrated that the abnormal prion protein (PrP^{SC}) responsible for vCJD (Will *et al*, 1996) was transmissible in blood. Human cases of vCJD and sub-clinical vCJD following earlier blood transfusion were later identified (Llewelyn *et al*, 2004; Peden *et al*, 2004).

As there is no test in blood for the abnormal prior protein (Collinge *et al*, 1996) that designates human exposure to Bovine Spongiform Encephalopathy (BSE) (Hill AF *et al*, 1997), the UK's blood supply was protected first by leucodepletion (that is: removal of leucocytes from blood donations); secondly, by barring those who have received blood or tissue from donating to others (Wroe *et al*, 2006; Clarke *et al*, 2007; Editorial, 2019).

Dietary exposure to BSE was extensive, albeit different by gender and birth-cohort, but has remained largely sub-clinical (Cooper *et al*, 2003). The presence of PrP^{SC} in lymphoid tissue such as spleen or appendix (Hilton *et al*, 1998; Hilton *et al*, 2004; Bishop *et al*, 2013, Gill *et al*, 2013) indicates exposure and subclinical carriage and aligns with BSE dietary exposure patterns; but does not imply progression to vCJD.

A1.5 Tuberculosis

Tuberculosis, a contagious bacterial disease, is one of the top 10 causes of death worldwide and is the leading infectious killer (WHO 2019 (A)). Co-infection with TB and HIV is a lethal combination that has been described as a synergy from hell (Bartlett, 2007). An estimated 1.5 million people died from TB in 2018, including 251,000 deaths amongst HIV-positive people (https://www.who.int/news-room/fact-sheets/detail/tuberculosis).

TB is widely regarded as a disease of poverty, although the disease affects individuals of all ages and socioeconomic status. **Active TB** commonly affects the lungs (pulmonary TB) but can affect other parts of the body (**extrapulmonary TB**).

About a quarter of the world's population has **latent TB infection** (LTBI), ie, they do not have symptoms of active TB and are not infectious. Nonetheless, people who have LTBI should be identified and treated because LTBI can develop into active TB; immunocompromised people, such as people living with HIV, malnutrition or diabetes, have a higher risk of developing active TB (https://www.who.int/news-room/fact-sheets/detail/tuberculosis). There are currently two methods to test for LTBI: the Mantoux tuberculin skin test (TST) and interferon gamma release assays (IGRAs). Neither of these tests can accurately **differentiate between TB infection and active TB disease**.

Early diagnosis of TB including universal **drug-susceptibility testing**, and systematic screening of contacts and high-risk groups is a component of the first of three pillars of the WHO's End TB Strategy. Sputum smear microscopy, developed more than 100 years ago, has been a diagnostic test for pulmonary TB particularly in low and middle income countries but has suboptimal performance and does not detect drug resistance. **Culture-based methods** are the reference standard. Culture yield varies with the severity of illness, specimen type and culture method (Nicol *et al*, 2011). Culture is generally regarded as an **imperfect reference standard** for TB detection and is frequently negative in extrapulmonary TB because of the paucibacillary (low bacterial load of *Mycobacterium tuberculosis*) nature of extrapulmonary specimens.

In the past 10 years, new **rapid molecular tests** that facilitate access to testing and produce results quicker than culture have been endorsed by the WHO for detection of active TB and drug resistance. In addition to assessing the overall performance of these tests in adults or children, assessments in **key subpopulations**, eg, according to HIV status and smear status, are necessary.

Diagnosis of **child TB disease** is frequently more challenging than in adults for two main reasons. First, **active TB in children** is typically paucibacillary. Therefore, even under ideal clinical and laboratory conditions, only 30% to 40% of child TB cases are microbiologically confirmed (Dunn *et al*, 2016). This is a major problem in assessing the performance of a new diagnostic test. Second, it is difficult to obtain sputum specimen from most children younger than six years old due to inability to expectorate. This is an important challenge since the **quality of a specimen** can affect the performance of a microbiological diagnostic test. For children who have difficulty producing sputum for detection of pulmonary TB, **alternative specimens** with different degrees of invasiveness, feasibility and acceptability, include induced sputum, gastric aspirate, nasopharyngeal aspirate and stool samples (because young children swallow their sputum). Due to the challenges of microbiological confirmation of TB in children, diagnosis of child tuberculosis relies on a mix of clinical, epidemiological, radiological, and laboratory information (Kay *et al*, 2020).

A1.6: Malaria

Malaria, like TB, is an ancient human disease and a major global public health challenge. **Five species of *Plasmodium* parasites**—*P. falciparum*, *P. vivax*, *P. malariae* and *P. ovale* (two species)—cause human malaria with the most severe form caused by *P. falciparum*. The parasites are transmitted to humans via mosquitoes

and the incubation period is 7 days or more. Individuals can be infected by multiple strains of the same species or by more than one *Plasmodium* species.

Malaria is an acute febrile illness with various **disease manifestations** in different patient populations and different epidemiologic settings, thus complicating diagnosis (Murphy *et al*, 2013). Immunity to *P. falciparum* malaria is acquired after years of repeated infections and wanes rapidly without ongoing parasite exposure (Weiss *et al*, 2010). Thus, in **malaria-endemic settings**, most adult malaria infections are subclinical and serve as reservoirs of infection for mosquitoes. Pregnant women and children are the two **most-at-risk groups** for malaria. The symptoms and complications of malaria in pregnancy depend on transmission setting and an individual's level of acquired immunity; in **high-transmission settings**, where levels of acquired immunity tend to be high, *P. falciparum* infection is usually asymptomatic in pregnancy (WHO 2017; WHO 2019b).

Presumptive treatment for malaria in febrile people in endemic settings facilitates overuse of antimalarials and development of drug resistance. Therefore, the WHO recommends that all cases of malaria should have a parasitological test (microscopy or RDT) to confirm the diagnosis (WHO 2015). **Malaria RDTs** detect parasite-specific antigens in the blood of infected individuals. Some RDTs detect only one species (*P. falciparum*), while other RDTs detect one or more of the other *Plasmodium* species. **Pan-specific RDTs** can distinguish *P. falciparum* (or mixed) infections from infections with only non-falciparum species. In people with malaria caused by *P. vivax*, relapses can occur because liver stages of the parasite can remain dormant and later cause symptomatic disease again. Thus, **vivax-specific RDTs** that can detect *P. vivax* from other *Plasmodium* species are important in *P. vivax* endemic regions. Similar to other biological tests, malaria RDTs can deteriorate when exposed to heat and humidity, and so need to be stable for them to be useful in malaria-endemic settings. The study design for the clinical/field evaluation of an RDT and the interpretation of results must take into account the likely **conditions of intended use** (Banoo *et al*, 2006).

High quality **microscopic examination** of thick and thin blood films is considered the **'gold standard'** for diagnosing malaria. In addition to parasite detection, microscopy allows differentiation of malaria species and parasite stages; determination of parasite density; assessment of drug effects; and can be used to diagnose other diseases (WHO 2016). However, the **accuracy and usefulness of microscopy** depends on the quality of the microscopes, reagents, experience of the microscopist, effective quality control and the quality assurance system (Ngasala *et al*, 2019). Pre-qualification of microscopists, blinded reading of blood films by more than one microscopist and a planned system to resolve discordant microscopy results are essential if microscopy is used as a gold standard (WHO 2009). Polymerase chain reaction (PCR) can be used to assess discordant results.

A1.7 Coronavirus disease (COVID-19)

Unlike SARS-CoV-1, which caused the SARS outbreak in 2002-2004, persons infected by SARS-CoV-2 are infectious for about two days before symptoms develop; and about one third of those infected may remain **asymptomatic** (Birrell, 2020). Although infected, infants and young children are less likely than secondary pupils, adults and senior citizens to progress to clinical disease. **Progression is strongly age-related**: During the UK's first wave, Birrell *et al* (2020) reporting on 29 October 2020, estimated the median infection fatality rate as 0.025% at 25-44 years (95% credible interval 0.018 to 0.033), 2.3% at 65-74 years (95% credible interval: 1.6% to 3.0%) but 16% at 75+ years (95% credible interval: 11% to 21%). Very sadly, during the first wave of the SARS-CoV-2 pandemic, half of those admitted to intensive care for COVID-19 disease died.

Unmitigated, the basic reproduction number for SARS-Cov-2 was extremely high, initially around 2.5 to 3, with a short doubling time of 3.3 days. The virus spreads efficiently between people through close contact indoors and via respiratory droplets from coughs and sneezes.

Since viral antigens are foreign proteins, people infected by SARS-CoV-2 generate an immune response and make **antibodies** that bind the nucleocapsid protein (N) and trimeric spike protein (S), and other viral proteins. The presence of antibodies specific for SARS-CoV-2 viral proteins therefore indicates that a person has been infected by the virus. Hence, tests that measure antibodies to N or S could be used in seroprevalence studies. Antibodies, such as IgG, that bind the receptor binding domain (RBD) of S—and thereby block the ability of the virus to attach to target cells—are most likely to neutralize virus infectivity. New variants of concern may escape detection by antibodies from previous SARS-CoV-2 infection or COVID-immunization.

Serum antibody levels peak at about one month after symptom-onset and then settle lower. Rate of decay of IgG antibody beyond 20 weeks is emerging science still (Gudbjartsson *et al*, 2020). High levels of neutralizing antibody may confer resistance to infection (To *et al*, 2020) or protect against severe disease (see also Young *et al*, 2020).

T cell responses are seen after infection with SARS-CoV-2 and T cell memory is important for sustained immunity against re-infection. The Karolinska COVID-19 Study Group has shown robust T cell immunity in convalescent individuals with asymptomatic or mild COVID-19 (Sekine *et al*, 2020). T cells help eliminate the virus from infected cells in the later stages of infection, but there is a downside: some T cells could be pathogenic (Sewell *et al*, 2020).

Appendix 2: Membership of the Working Party and Secretariat

Chairs

Professor Deborah Ashby OBE BSc MSc PhD CStat HonMFPHM HonMRCR FMedSci
Professor of Medical Statistics and Clinical Trials, Imperial College London.

Professor Jon Deeks BSc MSc PhD CStat FMedSci
Professor of Biostatistics, University of Birmingham.

Members

Professor Sheila Bird OBE MA PhD CStat FFPH FMedSci FRSE DSc(Hon.)
Formerly Programme Leader at MRC Biostatistics Unit, University of Cambridge.

Professor Stephen Evans BA MSc FRCP(Edin) Hon. FRCP(Lond)
Professor of Pharmacoepidemiology, London School of Hygiene and Tropical Medicine.

Professor Rafael Perera MA MSc DPhil
Professor of Medical Statistics, University of Oxford.

Professor Yemisi Takwoingi DVM MSc PhD
Professor of Test Evaluation and Evidence Synthesis, University of Birmingham.

Secretariat

Olivia Varley-Winter

Declaration of interests

Professor Deborah Ashby is a co-investigator on the REACT studies.

Professor Jon Deeks is a member of the MHRA IVD external advisory group; consultant to the WHO Essential Diagnostics List; lead of the Cochrane COVID-19 diagnostic test reviews team; co-investigator of the Birmingham University evaluation of the Innova test; lead of the NIHR Birmingham BRC Diagnostic and Biomarkers theme; chief methodological editor for Cochrane's diagnostic test accuracy systematic reviews; and co-applicant on MRC and NIHR grants on test evaluation. He is a member of the Royal Statistical Society's COVID-19 Taskforce and the RSS/DHSC Panel on NHS Test and Trace.

Professor Sheila M. Bird is a member of Royal Statistical Society's 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 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 Stephen Evans has no conflicts of interest.

Professor Rafael Perera 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). He is also deputy lead for the Multimorbidity and Long-term conditions Theme of the NIHR Oxford Biomedical Research Centre.

Professor Yemisi Takwoingi is a co-convenor of the Cochrane Screening and Diagnostic Tests Methods Group; Editor, Cochrane Infectious Diseases Group; Statistical Editor, Cochrane Bone, Joint and Muscle Trauma Group; Editor, Cochrane Diagnostic Test Accuracy Editorial Team; and co-investigator on MRC and NIHR funded test evaluation projects and MRC funded COVID-19 projects.

Acknowledgements

We are grateful to staff at the MHRA for providing information on current regulatory issues, and members of the Royal Statistical Society Covid-19 Taskforce and Stian Westlake for reviewing the report.

Appendix 3: Glossary

Accuracy: The term accuracy is used in multiple ways in test research. In analytical validity studies it describes the closeness of agreement of a measured quantity with the true quantity. In clinical and field evaluations it describes the ability of a test to correctly identify those with and without the condition, with sensitivity and specificity being described as measures of test accuracy. To minimise confusion this report uses the alternative phrase "test performance" for this second use.

Add-on test: A test used after an initial diagnostic test with the aim of improving the accuracy of the overall testing strategy by combining the results of a series of tests. Used to complement the initial test as it might have better accuracy (sensitivity or specificity) but might be deemed more costly or invasive than the initial one.

Analytical performance: Evaluation of a test in ideal laboratory conditions. Evidence to answer the question: Can the test reliably identify the analyte/measure of interest?

Analytical sensitivity: A measure of the change in a measurement in response to a change in the stimulus, typically evidence of the ability of tests to detect different concentrations of the pathogen.

Analytical specificity: Quantifies how likely the assay is to give false positive results due to cross reaction or interference with other medical conditions or substances.

Bayes or Bayesian updating: The computation of post-test probabilities from pre-test probabilities and likelihood ratios using Bayes' Theorem, and used to calculate predictive values of tests across different prevalence of infection.

Bias: Systematic observed difference between the measurements obtained from the new test and the 'true value', which is usually obtained from the reference standard. As with imprecision, bias could be related to the measurement level itself.

Blinding: In studies of tests, blinding refers to the practice of obtaining results of the index test without knowledge of the results of the reference standard, and vice versa, to ensure that the test and verification are independent.

Case-control study: See two-group study

Clinical agreement study: A study in which the reference standard is suspected to provide imperfect information regarding the presence/absence of the target condition in the study participants. The summary statistics (positive and negative percent agreement) reported reflect this uncertainty.

Clinical evaluation: See field study

Clinical impact: Evaluation of the test and the benefits and harms to patients. Evidence to answer the question: Does the test improve patient outcomes?

Clinical performance: Evaluation of the test in the setting and population for its intended use. Evidence to answer the question: How does the test perform during clinical evaluations in a real-life setting in a relevant population?

Clinical sensitivity: The proportion of true positives identified by the index test out of those identified as positive by the reference standard.

Clinical specificity: The proportion of true negatives identified by the index test out of those identified as negative by the reference standard.

Cochrane systematic reviews of diagnostic test accuracy: Evidence syntheses of diagnostic accuracy studies that are part of Cochrane and follow Cochrane methodological standards. In particular these syntheses: search for all available studies that evaluate the accuracy of a test for an intended use, review their quality (see QUADAS-2) in terms of risk of bias and concerns about applicability, and if appropriate combine their results using meta-analysis.

Coefficient of variation: A measure of relative precision, comparing the unexplained variability to the average value, typically based on a ratio.

Correlation of test errors: In the context of multiple tests, it is a measure of the chances of a second or later test being falsely positive or negative relative to previous test results being falsely positive or negative. These multiple tests could represent different tests or the same test taken over different time periods. Naïve Bayesian updating to work out the combined probability of these tests assumes this correlation to be zero (no correlation).

Diagnostic sensitivity: See Clinical sensitivity

Diagnostic specificity: See Clinical specificity

Discrepant analysis: Use of an additional reference test only in those where index tests and original reference standards disagree. As testing is selectively based on observed results this approach will generate a biased result.

Disease spectrum: A term used to describe the heterogeneity in those with the target condition, which may relate to the ability of tests to identify them.

Empirical evaluation: An evaluation method based on verifiable facts from evidence obtained by observation or experiment as opposed to just theory.

False negative: Observing a negative index test result in an individual or sample that has the target condition – a test error.

False positive: Observing a positive index test result in an individual or sample that does not have the target condition – a test error.

Field study: An evaluation of the performance of a test undertaken in a real-world setting aiming to produce an applicable estimate. The participants included are those in whom the test would be used if implemented in practice, and the test is undertaken and interpreted in the same way that would occur in practice. These studies are typically undertaken in clinical or community settings.

Head-to-head comparison: A study in which the performance of two or more tests are directly compared, either by undertaking multiple tests in the same individuals, or by comparing groups (preferably created through randomisation) that receive one or more of the tests being compared; all participants also receiving a reference standard diagnosis.

Imprecision: The level of variability observed for the test, usually in relation to repeated measurements carried out at the same time; here related to the concept of random error. In some cases, the imprecision is related to the measurement level – as when imprecision increases at increased levels of the analyte.

In vitro diagnostic (IVD): IVDs are tests done on samples such as fluids or tissue that have been taken from the human body.

Index test: the new test or existing test of interest to be evaluated to estimate its performance.

Intended use: The use for which a test is intended, normally according to the manufacturer of the test but re-purposing is also possible. A defined target population, stage in the natural history of the disease, and expected outcomes from the test results should be identified. EU regulation on diagnostic medical devices requires, as part of the device description and specifications, a statement of its intended use and intended users. The intended use should provide "sufficient information to enable the user to understand the medical context and to allow the intended user to make a correct interpretation of the results". The list of items suggested for inclusion about a diagnostic medical device are:

(i) what is to be detected and/or measured;
(ii) its function such as screening, monitoring, diagnosis or aid to diagnosis, prognosis, prediction, companion diagnostic;
(iii) the specific disorder, condition or risk factor of interest that it is intended to detect, define or differentiate;
(iv) whether it is automated or not;
(v) whether it is qualitative, semi-quantitative or quantitative;
(vi) the type of specimen(s) required;
(vii) where applicable, the testing population;
(viii) the intended user;
(ix) in addition, for companion diagnostics, the relevant target population and the associated medicinal product(s).

Intended role: Generally based on where, in the diagnostic pathway the test is meant to be used and to resolve what clinical question. Examples of this are: for triage, add-on, replacement, or as a new test.

Likelihood ratio: The ratio of the odds that a positive (for positive likelihood ratio) or negative (for negative likelihood ratio) result would be observed in individuals with the target disorder compared to individuals without the target disorder. Likelihood ratios are used in Bayesian updating.

Limit of blank (LoB): The highest apparent analyte concentration expected to be found when replicates of a sample containing no analyte are tested.

Limit of detection (LoD): The smallest concentration of a measurand that can reliably be detected, sometimes referred to as 'analytical sensitivity'.

Limit of quantification (LoQ): At low concentrations, the measurement error at the LoD may still be too high for reliable quantification (eg, if it has been agreed that a measurement must have a coefficient of variation = (Standard deviation/Mean)x100 less than a fixed value). LoQ is the value at or above the LoD and at which requirements for precision of a measurement are met; also called-"functional sensitivity".

Linkage: Use of multiple data sources combined to gather more complete information about individuals, such as test results with personal characteristics.

Mathematical model: A description of a system using assumptions and mathematical equations with the aim of making predictions. Mathematical models of testing predict how test use impacts on patient and population outcomes.

Natural history of disease: Expected progression of a disease process in an individual over time, usually assumed in the absence of treatment. There will be substantial variation between individuals and so natural history should be considered as general guidance.

Negative percent agreement: The proportion of negatives identified by the index test out of those identified as negative by the reference standard. Used instead of specificity when the reference standard is suspected to provide imperfect information regarding the disease status of the study participants.

Negative predictive value (NPV): The proportion of those with a negative index test result who have a negative result on the reference standard. Estimated from a prospective field (in context) study that consecutively or randomly recruited participants or using estimates of diagnostic sensitivity, diagnostic specificity, and potential prevalence in a target population.

New test: A test used to create/open a new testing pathway. No previous test in this particular setting/context is available.

Non-accuracy impacts: refers to all other factors that might impact on the performance of a test, such as human preferences or biases that could impact on the collection of the samples in the first place**.**

Patient outcome study: A study in which the impact of testing on patients is assessed, ideally comparing outcomes between groups randomised to receive different tests or test strategies. Outcomes may include the diagnoses made (diagnostic yield), the treatments used (therapeutic yield), and differences in patient relevant outcomes.

Performance: Characteristics that summarise the quality of a test. This report uses the term specifically to describe the agreement of test results with the reference standard, based on terms such as sensitivity and specificity.

Positive percent agreement: The proportion of positives identified by the index test out of those identified as positive by the reference standard. Used instead of sensitivity when the reference standard is suspected to provide imperfect information regarding the disease status of the study participants.

Positive predictive value (PPV): The proportion of those with a positive index test result who have a positive result on the reference standard. Estimated from a prospective field (in context) study that consecutively or randomly recruited participants or using estimates of diagnostic sensitivity, diagnostic specificity, and potential prevalence in a target population.

Precision: The closeness of agreement of repeated measures of the same sample under the same conditions.

Prevalence: The proportion of a group who have the target condition at a given time. In test studies it is important to distinguish population prevalence (the proportion in the whole proportion) from the prevalence in those being tested which is likely to be higher as testing is usually selected based on risk of the target condition.

Reference change value: The difference between repeated test results large enough to exclude the variation inherent to both measurements with a known degree of probability.

Reference standard: The test or tests used to classify individuals according to whether they do or do not have the target condition. The reference standard may be a single test, or a combination of tests and information, including information available subsequently. Reference standards need to make accurate classifications as they provide the assumed truth of health status against which index tests are compared.

Repeatability: The precision of measurements when repeated in the same conditions, such as by the same observer on the same assay at the same time. A measure of precision.

Replacement test: A test used instead of an existing test. A replacement test is expected to have similar or higher accuracy than the one to be replaced while improving/maintaining the level of cost or ease of use.

Reporting standard: A document that lists the essential details of the objectives, methods and results of a study to ensure that essential facts are communicated in a full and transparent manner.

Reproducibility: Measures how likely it is to obtain the same result when repeated tests are carried out on the same sample/individual under different conditions.

Screening: Testing of apparently healthy individuals for a target condition, or for risk factors for a target condition.

Selection bias: A bias created by choosing a non-representative group of samples or individuals for inclusion in a study.

Sensitivity: See clinical sensitivity.

Specificity: See clinical specificity.

Surveillance: Studies that assess the prevalence and incidence of the disease or health state of interest in a population, often over time.

Spiked sample: A sample prepared by adding a known quantity of a pathogen to a matrix (eg, saliva, serum, viral transport media) which is close or identical to that of the sample of interest.

Target condition: The target condition is the disease or health state that the test aims to detect. The reference standard is used to classify individuals according to whether they do or do not have the target condition.

Target product profile (TPP): The purpose of a TPP is to outline the desired 'profile' or characteristics of a target product that is aimed at a particular disease or diseases. TPPs state intended use, target populations and other desired attributes of products, including safety and performance-related characteristics.

Threshold: Results of numerical tests are classified as test positive or test negative according to whether they are above or below a numerical threshold. Altering the threshold affects the sensitivity and specificity of the test.

Test accuracy study: See test performance study. This report uses *test performance* in place of *test accuracy* to distinguish between analytical and clinical or field studies.

Test performance study: A clinical or field study in which individuals are tested by one or more index tests and a reference standard; the findings categorised as true positive, true negative, false positive and false negative; and estimates of test accuracy (sensitivity, specificity, positive and negative predictive values) calculated.

Timing: Refers to both when (eg, at the same time as reference, immediately after/before, hours/days later) and how long the test/result takes to be obtained. There can be significant variability in timing between types of tests with point-of-care tests (POCT) usually taking shorter time than laboratory tests, eg, elimination of the need to transport the sample to a laboratory, but typically at the expense of accuracy.

Triage test: A test used at the start or early on in the clinical pathway to determine if further testing should take place**.** A triage test is easier to carry out, less invasive, cheaper and/or has shorter time to obtain results than subsequent tests but is expected to be less accurate (either low sensitivity or specificity).

True negative: Observing a negative index test result in an individual or sample that does not have the target condition – a correct test result.

True positive: Observing a positive index test result in an individual or sample that does have the target condition – a correct test result.

Two-gate study: See two-group study.

Two-group study: A study that recruits two sets of individuals, those that are already known to have the target condition and those that are known not to have the target condition. Two-group studies exclude individuals where this is unclear, who are often most likely to give false negative or false positive, leading to overestimation of test performance.

Validity: The degree to which conclusions from a study are warranted taking account of the study methods and the representativeness of the study sample. Issues to do with study design, such as blinding and randomisation affect interval validity, issues to do with the representativeness of the individuals recruited and the delivery of the test affect external validity.

References

References (Section 1)

Centers for Disease Control and Prevention (CDC). Lesson 1: Introduction to Epidemiology (online) in *Principles of Epidemiology in Public Health Practice Third Edition An Introduction to Applied Epidemiology and Biostatistics*. 2006. https://www.cdc.gov/csels/dsepd/ss1978/lesson1/section9.html (Accessed 21st May 2021).

Machalaba CC, Loh EH, Daszak P, Karesh WB. Emerging Diseases from Animals. *State of the World 2015*. 2015;105-116. doi: 10.5822/978-1-61091-611-0_8.

van Seventer JM, Hochberg NS. Principles of Infectious Diseases: Transmission, Diagnosis, Prevention, and Control. International Encyclopedia of Public Health. 2017:22–39.

References (Section 2)

Banoo S, Bell D, Bossuyt P, Herring A, Mabey D, Poole F, Smith PG, Sriram N, Wongsrichanalai C, Linke R, O'Brien R, Perkins M, Cunningham J, Matsoso P, Nathanson CM, Olliaro P, Peeling RW, Ramsay A; TDR Diagnostics Evaluation Expert Panel (WHO/TDR) 2006. Evaluation of diagnostic tests for infectious diseases: general principles. *Nat Rev Microbiol*. 2006;4(9 Suppl):S21-31. doi: 10.1038/nrmicro1523.

Ehrmeyer SS, Laessig RH. Point-of-care testing, medical error, and patient safety: a 2007 assessment. *Clin Chem Lab Med.* 2007;45(6):766-73. doi: 10.1515/CCLM.2007.164.

References (Section 3)

Armbruster DA, Pry T. Limit of blank, limit of detection and limit of quantitation. *Clin Biochem Rev*. 2008;29 Suppl 1(Suppl 1):S49-S52.

Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet*. 1995;346(8982):1085-7. doi: 10.1016/s0140-6736(95)91748-9.

Carey RN, Anderson FP, George H. User demonstration of performance for precision and accuracy; approved guidelines–second edition. *CLSI document EP15-A2, Clinical and Laboratory Standard Institute.* 2005*;* 25:1-49.

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

Grijalva CG, Zhu Y, Halasa NB, Kim A, Rolfes MA, Steffens A, Reed C, Fry AM, Talbot H. High concordance between self-collected nasal swabs and saliva samples for detection of SARS-CoV-2. *Open Forum Infectious Diseases.* 2020: 7(S1),:S283. doi:10.1093/ofid/ofaa439.626.

Horvath AR, Lord SJ, StJohn A, Sandberg S, Cobbaert CM, Lorenz S, Monaghan PJ, Verhagen-Kamerbeek WD, Ebert C, Bossuyt PM; Test Evaluation Working Group of the European Federation of Clinical Chemistry

Laboratory Medicine. From biomarkers to medical tests: the changing landscape of test evaluation. *Clin Chim Acta*. 2014;427:49-57. doi: 10.1016/j.cca.2013.09.018.

Johnson R. Assessment of bias with emphasis on method comparison. *Clin Biochem Rev*. 2008;29 Suppl 1(Suppl 1):S37-42.

Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999;282(11):1061-6. doi: 10.1001/jama.282.11.1061.

Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. *Med Decis Making.* 2009;29(5):E13-21. doi: 10.1177/0272989X09336144.

Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case-control and two-gate designs in diagnostic accuracy studies. *Clin Chem.* 2005;51(8):1335-41. doi: 10.1373/clinchem.2005.048595.

Shinkins B, Thompson M, Mallett S, Perera R. Diagnostic accuracy studies: how to report and analyse inconclusive test results. *BMJ*. 2013;346:f2778. doi: 10.1136/bmj.f2778.

Takwoingi Y. Meta-analytic approaches for summarising and comparing the accuracy of medical tests. *University of Birmingham. Ph.D.* 2016. Available at: https://etheses.bham.ac.uk/id/eprint/6759/ [Accessed 19th May 2021].

Tholen D, Linnet K, Kondratovich M, Armbruster D, Garrett PE, Jones R, Kroll MH, Lequin R, Pankratz T, Scassellati GA, Schimmel H, Tsai J. Protocols for Determination of Limits of Detection and Limits of Quantitation; Approved Guidelines. *NCCLS document EP17-A, National Committee of Clinical and Laboratory Standards. 2004;*24: 1-39.

References (Section 4)

Alonzo TA, Pepe MS, Moskowitz CS. Sample Size Calculations for Comparative Studies of Medical Tests for Detecting Presence of Disease. *Stat Med*. 2002;21:835–52. doi: 10.1002/sim.1058.

Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: literature survey. *BMJ* 2006;332(7550):1127–9. doi: 10.1136/bmj.38793.637789.2F.

Bird AG, Gore SM, Jolliffe DW, Burns SM. Anonymous HIV surveillance in Saughton Prison, Edinburgh. *AIDS* 1992; 6: 725-733. doi: 10.1097/00002030-199207000-00017.

Bochmann F, Johnson Z, Azuara-Blanco A. Sample size in studies on diagnostic accuracy in ophthalmology: a literature survey. *Br J Ophthalmol* 2007;91(7):898–900. doi: 10.1136/bjo.2006.113290.

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

Doust JA, Bell KJL, Leeflang MMG, Dinnes J, Lord S J, Mallett S, van de Wijgert JHHM, Sandberg S, Adeli K, Deeks JJ, Bossuyt PM, Horvarth AR. Guidance for the design and reporting of studies evaluating the clinical performance of tests for present or past SARS-CoV-2 infection *BMJ* 2021;372:n568 doi: 10.1136/bmj.n568.

Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PMM, Deeks JJ. Assessing the value of diagnostic tests: a framework for designing and evaluating trials *BMJ* 2012; 344 :e686 doi: 10.1136/bmj.e686.

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

Gore SM, Bird AG, Burns SM, Goldberg DJ, Ross AJ, Macgregor J. Drug injection and HIV prevalence in inmates of Glenochil prison. *BMJ*. 1995;310(6975):293-6. doi: 10.1136/bmj.310.6975.293.

Gore SM, Bird AG, Cameron SO, Hutchinson SJ, Burns SM, Goldberg DJ. Prevalence of Hepatitis C carriage in Scottish prisons: WASH-C surveillance linked to self-reported risk behaviours. *QJM: An International Journal of Medicine* 1999;92(1):25-32. doi: 10.1093/qjmed/92.1.25.

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

Holtman GA, Berger MY, Burger H, Deeks JJ, Donner-Banzhoff N, Fanshawe TR, Koshiaris C, Leeflang MM, Oke JL, Perera R, Reitsma JB, Van den Bruel A. Development of practical recommendations for diagnostic accuracy studies in low prevalence situations. *J Clin Epidemiol*. 2019;114:38-48. doi: 10.1016/j.jclinepi.2019.05.018.

Horvath AR, Lord SJ, St John A, Sandberg S, Cobbaert CM, Lorenz S, Monaghan PJ, Verhagen-Kamerbeek WD, Ebert C, Bossuyt PM; Test Evaluation Working Group of the European Federation of Clinical Chemistry Laboratory Medicine. From biomarkers to medical tests: the changing landscape of test evaluation. *Clin Chim Acta*. 2014;427:49-57. doi: 10.1016/j.cca.2013.09.018.

Hutchinson SJ, Gore SM, Taylor A, Goldberg DJ, Frischer M. Extent and contributing factors of drug expenditure of injectors in Glasgow. Multi-site city-wide cross-sectional study. Br J Psychiatry. 2000;176:166-72. doi: 10.1192/bjp.176.2.166.

Iyer AS, Jones FK, Nodoushani A, Kelly M, Becker M, Slater D, Mills R, Teng E, Kamruzzaman M, Garcia-Beltran WF, Astudillo M, Yang D, Miller TE, Oliver E, Fischinger S, Atyeo C, Iafrate AJ, Calderwood SB, Lauer SA, Yu J, Li Z, Feldman J, Hauser BM, Caradonna TM, Branda JA, Turbett SE, LaRocque RC, Mellon G, Barouch DH, Schmidt AG, Azman AS, Alter G, Ryan ET, Harris JB, Charles RC. Persistence and decay of human antibody responses to the receptor binding domain of SARS-CoV-2 spike protein in COVID-19 patients. *Sci Immunol*. 2020;5(52):eabe0367. doi: 10.1126/sciimmunol.abe0367.

Korevaar DA, Gopalakrishna G, Cohen, JF, Bossuyt PM. Targeted test evaluation: a framework for designing diagnostic accuracy studies with clear study hypotheses. *Diagn Progn Res*. 2019;3,22. doi: 10.1186/s41512-019-0069-2.

Naaktgeboren CA, Bertens LCM, van Smeden M, de Groot JAH, Moons KGM, Reitsma JB. Value of composite reference standards in diagnostic research *BMJ*. 2013.347:f5605. doi: 10.1136/bmj.f5605.

Obuchowski NA. Sample size calculations in studies of test accuracy. *Stat Methods Med Res*. 1998;7(4):371-92. doi: 10.1177/096228029800700405.

Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;299(17):926-30. doi: 10.1056/NEJM197810262991705.

Thombs BD, Rice DB. 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. *Int J Methods Psychiatr Res* 2016;25(2):145–52. doi:10.1002/mpr.1504.

White SR, Hutchinson SJ, Taylor A, Bird SM. Modeling the initiation of others into injection drug use, using data from 2,500 injectors surveyed in Scotland during 2008-2009. *Am J Epidemiol*. 2015;181(10):771-780. doi: 10.1093/aje/kwu345.

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

Yirrell DL, Robertson P, Goldberg DJ, McMenamin J, Cameron S, Leigh Brown AJ. Molecular investigation into outbreak of HIV in a Scottish prison. *BMJ*. 1997;314(7092):1446-50. doi: 10.1136/bmj.314.7092.1446.

References (Section 5)

Abbott Diagnostics. ID Now Instructions for Use. IN 190000 Rev. 7 2020/09. Available at: https://www.fda.gov/media/136525/download. [Accessed 21st May 2021].

Abingdon Health (A). UK COVID-19 rapid antibody tests approved for professional use. Available at: https://www.abingdonhealth.com/uk-COVID-19-rapid-antibody-tests-approved-for-professional-use/ [Accessed 19th May 2021].

Abingdon Health (B). ABC-19 response to media. Available at: https://www.abingdonhealth.com/news/abc-19-response-to-media/ [Accessed 21st May 2021].

Adams ER, Ainsworth M, Anand R, Andersson MI, Auckland K, Baillie JK, Barnes E, Beer S, Bell JI, Berry T, Bibi S, Carroll M, Chinnakannan SK, Clutterbuck E, Cornall RJ, Crook DW, de Silva T, Dejnirattisai W, Dingle KE, Dold C, Espinosa A, Eyre DW, Farmer H, Fernandez Medoza M, Georgiou D, Hoosdally SJ, Hunter A, Jefferey K, Kelly DF, Klenerman P, Knight J, Knowles C, Kwok AJ, Leuschner U, Levin R, Liu C, López-Camacho C, Martinez J, Matthews PC, McGivern H, Mentzer AJ, Milton J, Mongkolsapaya J, Moore SC, Oliveira MS, Pereira F, Perez E, Peto T, Ploeg RJ, Pollard A, Prince T, Roberts DJ, Rudkin JK, Sanchez V, Screaton GR, Semple MG, Slon-Campos J, Skelly DT, Smith EN, Sobrinodiaz A, Staves J, Stuart DI, Supasa P, Surik T, Thraves H, Tsang P, Turtle L, Walker AS, Wang B, Washington C, Watkins N, Whitehouse J, National COVID Testing Scientific Advisory Panel. Antibody testing for COVID-19: A report from the National COVID Scientific Advisory Panel [version 1; peer review: 2 approved]. *Wellcome Open Res.* 2020.5:139. doi: 10.12688/wellcomeopenres.15927.1.

Allan C, Joyce TJ, Pollock AM. Europe's new device regulations fail to protect the public. *BMJ*. 2018;363:k4205. doi: 10.1136/bmj.k4205.

Arevalo-Rodriguez I, Buitrago-Garcia D, Simancas-Racines D, Zambrano-Achig P, Del Campo R, Ciapponi A, Sued O , Martinez-Garcıa L, Rutjes AW, Low N, Bossuyt PM, Perez-Molina JA, Zamora J. False-negative results of initial RT-PCR assays for COVID-19: A systematic review. *PLoS ONE* 2020.15(12):e0242958. doi: 10.1371/journal.pone.0242958.

Boseley S. UK coronavirus home testing to be made available to millions. *The Guardian* 25th March 2020. Available at: https://www.theguardian.com/world/2020/mar/25/uk-coronavirus-mass-home-testing-to-be-made-available-within-days. [Accessed 17th May 2021].

Buchanan L, Gamio L, Leatherby L, Keefe J, Koettl C, Walker AS. Tracking the White House Coronavirus Outbreak. *The New York Times* 2 October, 2020. Available at: https://www.nytimes.com/interactive/2020/10/02/us/politics/trump-contact-tracing-covid.html [Accessed October 5, 2020].

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

Cao, S., Gan, Y., Wang, C. *et al*. Post-lockdown SARS-CoV-2 nucleic acid screening in nearly ten million residents of Wuhan, China. *Nat Commun.* 2020.11;5917. doi: 10.1038/s41467-020-19802-w.

Case JB, Bailey AL, Kin AS, Chen RE, Diamond MS. Growth, detection, quantification and inactivation of SARS-CoV-2. *Virology.* 2020.548:39-48. doi: 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. Performance of VivaDiag COVID-19 IgM/IgG Rapid Test is inadequate for diagnosis of COVID-19 in acute patients referring to emergency room department. J Med Virol. 2020;92(10):1724-1727. doi: 10.1002/jmv.25800.

Cevik M, Tate M, Lloyd O, Maraolo AE, Schafers J, Ho A. SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis. Lancet Microbe 2021; 2(1): e13–22. doi: 10.1016/ S2666-5247(20)30172-5.

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

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

Deeks JJ, Raffle A, Gill M. 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 January 29th 2021. Available at: https://www.bmj.com/content/372/bmj.n81/rapid-responses [Accessed 19th May 2021].

Deeks J, Gill M, Bird S, Richardson S, Ashby D. COVID-19 Innova testing in schools: don't just test, evaluate. BMJ January 12th 2021. Available at: https://blogs.bmj.com/bmj/2021/01/12/covid-19-innova-testing-in-schools-dont-just-test-evaluate/ [Accessed 19th May 2021].

Deeks J. Why the school testing regimen needs to change. *UnHerd* 10th March 2021. Available at: https://unherd.com/thepost/why-the-school-testing-regime-needs-to-change/ [Accessed 19th May 2021].

Department of Health and Social Care (A). Rapid evaluation of OptiGene RT-LAMP assay (direct and RNA formats). 1st December 2020. Available at: https://www.gov.uk/government/publications/rapid-evaluation-of-optigene-rt-lamp-assay-direct-and-rna-formats [Accessed 19th May 2021].

Department of Health and Social Care (B). Press release. Government invests in UK-developed antibody tests from UK Rapid Test Consortium. 6 October 2020. Available at: https://www.gov.uk/government/news/government-invests-in-uk-developed-antibody-tests-from-uk-rapid-test-consortium [Accessed 19th May 2021].

Department of Health and Social Care (C). 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. Available at: https://www.gov.uk/government/publications/weekly-statistics-for-nhs-test-and-trace-england-15-april-to-21-april-2021 [Accessed 3rd May 2021].

Diggle PJ. Estimating prevalence using an imperfect test. *Epidemiology Research International 2011*; 2011;608719. doi: 10.1155/2011/608719.

Dinnes J, Deeks JJ, 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 IM, Price MJ, Taylor-Phillips S, Hooft L, Leeflang MMG, McInnes MDF, Spijker R, Van den Bruel A. Rapid, point-of-care antigen and molecular-based tests for diagnosis of SARS-CoV-2 infection. Cochrane Database of Systematic Reviews 2021, Issue 3. Art. No.: CD013705. DOI: 10.1002/14651858.CD013705.pub2. [Accessed 06 April 2021].

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. 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 Biol*. 2021;19(4):e3001216. doi: 10.1371/journal.pbio.3001216.

Flower B, Brown JC, Simmons B, Moshe M, Frise R, Penn R, Kugathasan R, Petersen C, Daunt A, Ashby D, Riley S, Atchison CJ, Taylor GP, Satkunarajah S, Naar L, Klaber R, Badhan A, Rosadas C, Khan M, Fernandez N, Sureda-Vives M, Cheeseman HM, O'Hara J, Fontana G, Pallett SJC, Rayment M, Jones R, Moore LSP, McClure MO, Cherepanov P, Tedder R, Ashrafian H, Shattock R, Ward H, Darzi A, Elliot P, Barclay WS, Cooke GS. Clinical and laboratory evaluation of SARS-CoV-2 lateral flow assays for use in a national COVID-19 seroprevalence survey. Thorax. 2020;75(12):1082-1088. doi: 10.1136/thoraxjnl-2020-215732.

Foundation for Innovative New Diagnostics. SARS-CoV-2 diagnostic use cases. Available at: https://www.finddx.org/COVID-19-old/dx-use-cases/. [Accessed 12th April 2021].

Gudbjartsson DF, Norddahl GL, Melsted P, Gunnarsdottir K, Holm H, Eythorsson E, Arnthorsson AO, Helgason D, Bjarnadottir K, Ingvarsson RF, Thorsteinsdottir B, Kristjansdottir S, Birgisdottir K, Kristinsdottir AM, Sigurdsson MI, Arnadottir GA, Ivarsdottir EV, Andresdottir M, Jonsson F, Agustsdottir AB, Berglund J, Eiriksdottir B, Fridriksdottir R, Gardarsdottir EE, Gottfredsson M, Gretarsdottir OS, Gudmundsdottir S, Gudmundsson KR, Gunnarsdottir TR, Gylfason A, Helgason A, Jensson BO, Jonasdottir A, Jonsson H, Kristjansson T, Kristinsson KG, Magnusdottir DN, Magnusson OT, Olafsdottir LB, Rognvaldsson S, le Roux L, Sigmundsdottir G, Sigurdsson A, Sveinbjornsson G, Sveinsdottir KE, Sveinsdottir M, Thorarensen EA, Thorbjornsson B, Thordardottir M, Saemundsdottir J, Kristjansson SH, Josefsdottir KS, Masson G, Georgsson G, Kristjansson M, Moller A, Palsson R, Gudnason T, Thorsteinsdottir U, Jonsdottir I, Sulem P, Stefansson K. Humoral Immune Response to SARS-CoV-2 in Iceland. *N Engl J Med*. 2020;383(18):1724-1734. doi: 10.1056/NEJMoa2026116.

Haan L, Ferreira A. Extreme value theory: an introduction. Springer 2007.

Hall VJ, Foulkes S, Charlett A, Atti A, Monk EJM, Simmons R, Wellington E, Cole MJ, Saei A, Oguti B, Munro K, Wallace S, Kirwan PD, Shrotri M, Vusirikala A, Rokadiya S, Kall M, Zambon M, Ramsay M, Brooks T, Brown CS, Chand MA, Hopkins S; SIREN Study Group. 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*. 2021;397(10283):1459-1469. doi: 10.1016/S0140-6736(21)00675-9.

Holtman GA, Berger MY, Burger H, Deeks JJ, Donner-Banzhoff N, Fanshawe TR, Koshiaris C, Leeflang M, Oke J, Perera R, Reitsma J, Van den Bruel A. Development of practical recommendations for diagnostic accuracy studies in low-prevalence situations. *J Clin Epidemiol*. 2019;114:38-48. doi: 10.1016/j.jclinepi.2019.05.018.

Hung IF, Cheng VC, Li X, Tam AR, Hung DL, Chiu KH, Yip CC, Cai JP, Ho DT, Wong SC, Leung SS, Chu MY, Tang MO, Chen JH, Poon RW, Fung AY, Zhang RR, Yan EY, Chen LL, Choi CY, Leung KH, Chung TW, Lam SH, Lam TP, Chan JF, Chan KH, Wu TC, Ho PL, Chan JW, Lau CS, To KK, Yuen KY. SARS-CoV-2 shedding and seroconversion among passengers quarantined after disembarking a cruise ship: a case series. *Lancet Infect Dis*. 2020;20(9):1051-1060. doi: 10.1016/S1473-3099(20)30364-9.

Innova Medical Group (A). SARS-CoV-2 antigen rapid qualitative test. Instructions for use. Version A/02 2020-07-01. Available at: https://cdn.website-editor.net/6f54caea7c6f4adfba8399428f3c0b0c/files/uploaded/Innova-SARS-Cov-2-Antigen-test-IFU.pdf [Accessed 19th May 2021].

Innova Medical Group (B). Innova rapid antigen test as a public health screening tool. https://innovamedgroup.com/innova-rapid-antigen-test-as-a-public-health-screening-tool/ [Accessed 28th April 2021].

ISO (International Organisation for Standardization) In vitro diagnostic medical devices – Clinical performance studies using specimens from human subjects – Good study practice. ISO 20916:2019 Available at: https://standards.iteh.ai/catalog/standards/sist/a4770d92-b282-49e5-b2e1- c719ca76aa1b/iso-20916-2019. [Accessed 19th April 2021].

Larremore DB, Wilder B, Lester E, Shehata S, Burke JM, Hay JA, Tambe M, Mina MJ, Parker R. Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening. *Sci Adv.* 2021;7(1):eabd5393. doi: 10.1126/sciadv.abd5393.

Lord SJ, Staub LP, Bossuyt PMM, Irwig LM. Target practice: choosing target conditions for test accuracy studies that are relevant to clinical practice. *BMJ.* 2011;343:d4684 doi:10.1136/bmj.d4684.

Mandavilli A. The White House relied on a rapid test, but used it in a way it was not intended. The New York Times, 2nd October 2020. Available at: 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 [Accessed 18th May 2021].

Medicines and Healthcare products Regulatory Agency (MHRA). New story: action taken to halt sales of fingerprick coronavirus (COVID-19) antibody testing kits. 29th May 2020. Available at: https://www.gov.uk/government/news/action-taken-to-halt-sales-of-fingerprick-coronavirus-COVID-19-antibody-testing-kits [Accessed 8th April 2021].

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

Mitjà O, Corbacho-Monné M, Ubals M, Alemany A, Suñer C, Tebé C, Tobias A, Peñafiel J, Ballana E, Pérez CA, Admella P, Riera-Martí N, Laporte P, Mitjà J, Clua M, Bertran L, Sarquella M, Gavilán S, Ara J, Argimon JM, Cuatrecasas G, Cañadas P, Elizalde-Torrent A, Fabregat R, Farré M, Forcada A, Flores-Mateo G, López C, Muntada E, Nadal N, Narejos S, Nieto A, Prat N, Puig J, Quiñones C, Ramírez-Viaplana F, Reyes-Urueña J, Riveira-Muñoz E, Ruiz L, Sanz S, Sentís A, Sierra A, Velasco C, Vivanco-Hidalgo RM, Zamora J, Casabona J, Vall-Mayans M, González-Beiras C, Clotet B; BCN-PEP-CoV2 Research Group. A Cluster-Randomized Trial of

Hydroxychloroquine for Prevention of Covid-19. *N Engl J Med*. 2021;384(5):417-427. doi: 10.1056/NEJMoa2021801.

Mizumoto K, Kagaya K, Zarebski A, Chowell G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Euro Surveill.* 2020;25(10):pii=2000180. doi: 10.2807/1560- 7917.ES.2020.25.10.2000180.

Moshe M, Daunt A, Flower B, Simmons B, Brown JC, Frise R, Penn R, Kugathasan R, Petersen C, Stockmann H, Ashby D, Riley S, Atchison C, Taylor GP, Satkunarajah S, Naar L, Klaber R, Badhan A, Rosadas C, Marchesin F, Fernandez N, Sureda-Vives M, Cheeseman H, O'Hara J, Shattock R, Fontana G, Pallett SJC, Rayment M, Jones R, Moore LSP, Ashrafian H, Cherapanov P, Tedder R, McClure M, Ward H, Darzi A, Elliott P, Cooke GS, Barclay WS; React study team. SARS-CoV-2 lateral flow assays for possible use in national covid-19 seroprevalence surveys (React 2): diagnostic accuracy study. *BMJ*. 2021;372:n423. doi: 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 AE, Wyllie DH on behalf of the EDSAB-HOME and COMPARE Investigators. 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.* 2020;371:m4262 doi: 10.1136/bmj.m4262.

NHS Test and Trace. COVID-19 National Testing Programme: Schools & Colleges handbook. 15 Dec 2020. Available at: https://schoolsweek.co.uk/wp-content/uploads/2020/12/Schools_Colleges_Testing-Handbook_version-3.3-Copy.pdf. [Accessed 19th May 2021].

Office for National Statistics. Coronavirus (COVID-19) Infection Survey, UK: 21 August 2020. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveypilot/englandandwales21august2020. [Accessed 19th May 2021].

Our World in Data. Australia: Coronavirus Pandemic Country Profile. Available at: https://ourworldindata.org/coronavirus/country/australia [Accessed 1st May 2021].

Peto T, UK COVID-19 Lateral Flow Oversight Team. COVID-19: Rapid Antigen detection for SARS-CoV-2 by lateral flow assay: a national systematic evaluation for mass-testing *medRxiv* [preprint] 2021.01.13.21249563; doi: 10.1101/2021.01.13.21249563.

Pickering S, Batra R, Snell LB, Merrick B, Nebbia G, Douthwaite S, Patel A, Ik MTK, Patel B, Charalampous T, Alcolea-Medina A, Lista MJ, Cliff PR, Cunningham E, Mullen J, Doores KJ, Edgeworth JD, Malim MH, Neil SJD, Galão RP. Comparative performance of SARS CoV-2 lateral flow antigen tests demonstrates their utility for high sensitivity detection of infectious virus in clinical specimens. *medRxiv* [preprint] 2021.02.27.21252427; doi: 10.1101/2021.02.27.21252427.

Post N, Eddy D, Huntley C, van Schalkwyk MCI, Shrotri M, Leeman D, Rigby S, Williams SV, Bermingham WH, Kellam P, Maher J, Shields AM, Amirthalingam G, Peacock SJ, Ismail SA. Antibody response to SARS-CoV-2 infection in humans: A systematic review. *PLoS One*. 2020;15(12):e0244126. doi: 10.1371/journal.pone.0244126.

Pray IW, Ford L, Cole D, Lee C, Bigouette JP, Abedi GR, Bushman D, Delahoy MJ, Currie D, Cherney B, Kirby M, Fajardo G, Caudill M, Langolf K, Kahrs J, Kelly P, Pitts C, Lim A, Aulik N, Tamin A, Harcourt JL, Queen K, Zhang J, Whitaker B, Browne H, Medrzycki M, Shewmaker P, Folster J, Bankamp B, Bowen MD, Thornburg NJ, Goffard K, Limbago B, Bateman A, Tate JE, Gieryn D, Kirking HL, Westergaard R, Killerby M; CDC COVID-19 Surge Laboratory Group. Performance of an Antigen-Based Test for Asymptomatic and Symptomatic SARS-CoV-2

Testing at Two University Campuses - Wisconsin, September-October 2020. *MMWR Morb Mortal Wkly Rep*. 2021;69(5152):1642-1647. doi: 10.15585/mmwr.mm695152a3.

Public Health England. Evaluation of sensitivity and specificity of four commercially available SARS-CoV-2 antibody immunoassays. July 2020. Available at: 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. [Accessed 12 April 2021].

Quilty BJ, Clifford S, Hellewell J, Russell TW, Kucharski AJ, Flasche S, Edmunds WJ; Centre for the Mathematical Modelling of Infectious Diseases COVID-19 working group. Quarantine and testing strategies in contact tracing for SARS-CoV-2: a modelling study. Lancet Public Health. 2021;6(3):e175-e183. doi: 10.1016/S2468-2667(20)30308-X. Epub 2021 Jan 21. PMID: 33484644; PMCID: PMC7826085.

Ramdas K, Darzi A, Jain S. 'Test, re-test, re-test': using inaccurate tests to greatly increase the accuracy of COVID-19 testing. *Nat Med* 2020;26:810–811. doi: 10.1038/s41591-020-0891-7.

REACT Study Investigators, Riley S, Ainslie KE, Eales O, Walters CE, Wang H, Atchison C, Diggle PJ, Ashby D, Donnelly CA, Cooke G, Barclay W, Ward H, Darzi A, Elliott P. Transient dynamics of SARS-CoV-2 as England exited national lockdown. *medRxiv* [preprint] 2020.08.05.20169078. doi: 10.1101/2020.08.05.20169078.

Robertson LJ, Moore JS, Blighe K, Ng KY, Quinn N, Jennings F, Warnock G, Sharpe P, Clarke M, Maguire K, Rainey S, Price R, Burns W, Kowalczyk A, Awuah A, McNamee S, Wallace G, Hunter D, Segar S, Shern CC, Nesbit MA, McLaughlin J, Moore T. Laboratory evaluation of SARS-CoV-2 antibodies: detectable IgG up to 20 weeks post infection. *medRxiv* [preprint] 2020.09.29.20201509; doi: https://doi.org/10.1101/2020.09.29.20201509.

Royal Society of Medicine COVID-19 series: challenges of testing – Episode 38. Professor Sir John Bell [Interview]. Available at: https://www.rsm.ac.uk/events/rsm-studios/2019-20/pen97/ [Accessed 19th May 2021].

Royal Statistical Society 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, 5 March 2021. Available at: https://rss.org.uk/RSS/media/File-library/News/2021/RSS-statement-on-surveillance-in-schools-5-March-2021.pdf [Accessed 19th May 2021].

Scientific Advisory Group for Emergencies. Minutes Meeting 73, 17th December. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/952613/s0989-covid-19-sage-73-minutes-171220.pdf [Accessed 19th May 2021].

Schuit E, Veldhuijzen IK, Venekamp RP, van den Bijllaardt W, Pas SD, Lodder EB, Molenkamp R, GeurtsvanKessel CH, Velzing J, Huisman RC, Brouwer L, Boelsums T, Sips GJ, Benschop KSM, Hooft L, van de Wijgert JHHM , van den Hof S , Moons KGM. Diagnostic accuracy of rapid antigen tests in pre-/asymptomatic close contacts of individuals with a confirmed SARS-CoV-2 infection. *medRxiv* [preprint] 2021.03.18.21253874. doi: https://doi.org/10.1101/2021.03.18.21253874.

Shuren J, Stenzel T. Covid-19 Molecular Diagnostic Testing - Lessons Learned. *N Engl J Med*. 2020 Oct 22;383(17):e97. doi: 10.1056/NEJMp2023830.

Shuren J, Stenzel T. The FDA's experience with COVID-19 antibody tests. *N Engl J Med* 2021; 384:592-594 doi: 10.1056/NEJMp2033687.

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

Takwoingi Y, Leeflang MM, Deeks JJ. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. *Ann Intern Med*. 2013;158(7):544-54. doi: 10.7326/0003-4819-158-7-201304020-00006.

University of Liverpool. Liverpool COVID-19 community testing pilot. Interim evaluation report. Available at: https://www.liverpool.ac.uk/media/livacuk/coronavirus/Liverpool,Community,Testing,Pilot,Interim,Evaluation.pdf [Accessed 19th May 2021].

Ward H, Cooke G, Atchison C, Whitaker M, Elliott J, Moshe M, Brown JC, Flower B, Daunt A, Ainslie K, Ashby D, Donnelly C, Riley S, Darzi A, Barclay W, Elliott P, for the REACT study team. Declining prevalence of antibody positivity to SARS-CoV-2: a community study of 365,000 adults. *medRxiv* [preprint] 2020.10.26.20219725; doi: 10.1101/2020.10.26.20219725.

Ward H, Cooke G, Whitaker M, Redd R, Eales O, Brown JC, Collet K, Cooper E, Daunt A, Jones K, Moshe M, Willicombe M, Day S, Atchison C, Darzi A, Donnelly CA, Riley S, Ashby D, Barclay WS, Elliott P. 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* [preprint] 2021.02.26.21252512; doi: 10.1101/2021.02.26.21252512.

World Health Organization. 2020. Global surveillance for COVID-19 caused by human infection with COVID-19 virus: Interim guidance, 20th March 2020. Available at: https://www.who.int/publications/i/item/who-2019-nCoV-surveillanceguidance-2020.8 [accessed: 21 May 2021].

References (Section 6)

Medicines and Healthcare products Regulatory Agency. Decision: Medical devices given exceptional use authorisations during the COVID-19 pandemic. Available at: https://www.gov.uk/government/publications/medical-devices-given-exceptional-use-authorisations-during-the-covid-19-pandemic. [Accessed 1st May 2021].

Medicines and Healthcare products Regulatory Agency. Corporate report: MHRA Corporate Plan 2018 to 2023. Available at: https://www.gov.uk/government/publications/mhra-corporate-plan-2018-to-2023. [Accessed 1st May 2021].

European Parliament and Council. Directive 98/79/EC of the European Parliament and of the Council of 27 October 1998 on in vitro diagnostic medical devices. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:31998L0079. [Accessed 1st May 2021].

European Parliament and Council. 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. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0746&qid=1620089932271. [Accessed 1st May 2021].

References (Section 7)

Adams ER, Ainsworth M, Anand R, Andersson MI, Auckland K, Baillie JK, Barnes E, Beer S, Bell JI, Berry T, Bibi S, Carroll M, Chinnakannan SK, Clutterbuck E, Cornall RJ, Crook DW, de Silva T, Dejnirattisai W, Dingle KE,

Dold C, Espinosa A, Eyre DW, Farmer H, Fernandez Medoza M, Georgiou D, Hoosdally SJ, Hunter A, Jefferey K, Kelly DF, Klenerman P, Knight J, Knowles C, Kwok AJ, Leuschner U, Levin R, Liu C, López-Camacho C, Martinez J, Matthews PC, McGivern H, Mentzer AJ, Milton J, Mongkolsapaya J, Moore SC, Oliveira MS, Pereira F, Perez E, Peto T, Ploeg RJ, Pollard A, Prince T, Roberts DJ, Rudkin JK, Sanchez V, Screaton GR, Semple MG, Slon-Campos J, Skelly DT, Smith EN, Sobrinodiaz A, Staves J, Stuart DI, Supasa P, Surik T, Thraves H, Tsang P, Turtle L, Walker AS, Wang B, Washington C, Watkins N, Whitehouse J, National COVID Testing Scientific Advisory Panel. Antibody testing for COVID-19: A report from the National COVID Scientific Advisory Panel [version 1; peer review: 2 approved]. *Wellcome Open Res.* 2020.5:139. doi: 10.12688/wellcomeopenres.15927.1.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG Moher D, Rennie D, de Vet HCW, Kressel HY, Rifai N, Golub RM, Altman DG, Hooft L, Korevaar DA, Cohen JF, For the STARD Group. STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. *BMJ.* 2015;351:h5527. doi: 10.1136/bmj.h5527.

Chan AW, Hróbjartsson A. Promoting public access to clinical trial protocols: challenges and recommendations. *Trials*. 2018;19(1):116. doi:10.1186/s13063-018-2510-1.

Department of Health. Government response to the House of Commons Science and Technology Committee Report on National Health Screening. January 2015. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/399296/national_health_screening_8999.pdf [Accessed 19th May 2021].

Dwan K, Gamble C, Williamson PR, Kirkham JJ, 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. Doi:10.1371/journal.pone.0066844.

Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, Michie S, Moher D, Wager E. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet*. 2014;383(9913):267-76. doi: 10.1016/S0140-6736(13)62228-X.

Goodacre S, Irving A, Wilson R, Beever D, Challen K. The PAndemic INfluenza Triage in the Emergency Department (PAINTED) pilot cohort study. *Health Technol Assess* 2015;19(3). doi: 10.3310/hta19030.

McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM; and the PRISMA-DTA Group, Clifford T, Cohen JF, Deeks JJ, Gatsonis C, Hooft L, Hunt HA, Hyde CJ, Korevaar DA, Leeflang MMG, Macaskill P, Reitsma JB, Rodin R, Rutjes AWS, Salameh JP, Stevens A, Takwoingi Y, Tonelli M, Weeks L, Whiting P, Willis BH. Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies: The PRISMA-DTA Statement. *JAMA*. 2018;319(4):388-396. doi: 10.1001/jama.2017.19163

Salameh JP, Bossuyt PM, McGrath TA, Thombs BD, Hyde CJ, Macaskill P, Deeks JJ, Leeflang M, Korevaar DA, Whiting P, Takwoingi Y, Reitsma JB, Cohen JF, Frank RA, Hunt HA, Hooft L, Rutjes AWS, Willis BH, Gatsonis C, Levis B, Moher D, McInnes MDF. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. *BMJ*. 2020 Aug 14;370:m2632. doi: 10.1136/bmj.m2632.

Science and Technology Committee – 4 Communicating the risks and benefits of health screening. National Health Screening 2014.Point 52. Available at: https://publications.parliament.uk/pa/cm201415/cmselect/cmsctech/244/24407.htm#a13. [Accessed 19th May 2021].

Simes RJ. Publication bias: the case for an international registry of clinical trials. J Clin Oncol. 1986;4(10):1529-41. doi: 10.1200/JCO.1986.4.10.1529.

Taichman DB, Sahni P, Pinborg A, Peiperl L, Laine C, James A, Hong ST, Haileamlak A, Gollogly L, Godlee F, Frizelle FA, Florenzano F, Drazen JM, Bauchner H, Baethge C, Backus J. Data Sharing Statements for Clinical Trials: A Requirement of the International Committee of Medical Journal Editors. Ethiop J Health Sci. 2017;27(4):315-318.

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

References (Appendix 1)

Alexander TS. Human Immunodeficiency Virus Diagnostic Testing: 30 Years of Evolution. *Clin Vaccine Immunol*. 2016;23(4):249-53. doi: 10.1128/CVI.00053-16.

Babiker A, Darbyshire J, Pezzotti P, Porter K, Rezza G, Walker SA, Beral V, Coutinho R, Del Amo J, Gill N, Lee C, Meyer L, Tyrer F, Dabis F, Thiebaut R, Lawson-Aye S, Boufassa F, Hamouda O, Fischer K, Pezzotti P, Rezza G, Touloumi G, Hatzakis A, Karafoulidou A, Katsarou O, Brettle R, del Romero J, Prins M, van Benthem B, Kirk O, Pederson C, Hernández Aguado I, Pérez-Hoyos S, Eskild A, Bruun JN, Sannes M, Sabin C, Lee C, Johnson AM, Phillips AN, Francioli P, Vanhems P, Egger M, Rickenbach M, Cooper D, Kaldor J, Ashton L, Vizzard J, Muga R, Day NE, De Angelis D; CASCADE Collaboration. Changes over calendar time in the risk of specific first AIDS-defining events following HIV seroconversion, adjusting for competing risks. *Int J Epidemiol*. 2002;31(5):951-8. doi: 10.1093/ije/31.5.951.

Banoo S, Bell D, Bossuyt P, Herring A, Mabey D, Poole F, Smith PG, Sriram N, Wongsrichanalai C, Linke R, O'Brien R, Perkins M, Cunningham J, Matsoso P, Nathanson CM, Olliaro P, Peeling RW, Ramsay A; TDR Diagnostics Evaluation Expert Panel (WHO/TDR) 2006. Evaluation of diagnostic tests for infectious diseases: general principles. *Nat Rev Microbiol*. 2006 Sep;4(9 Suppl):S21-31. doi: 10.1038/nrmicro1523.

Bartlett JG, Tuberculosis and HIV Infection: Partners in Human Tragedy. *Journal of Infectious Diseases* 2007;196(1):S124–S125. doi: 10.1086/518668.

Bird SM. Like-with-like comparisons? *Lancet*. 2010;376(9742):684. doi: 10.1016/S0140-6736(10)61333-5.

Birrell PJ, Zhang X-S, Pebody RG, Gay NJ, De Angelis D. Reconstructing a spatially heterogeneous epidemic; characterising the geographic spread of 2009 A/H1N1pdm infection in England. *Scientific Reports.* 2016;6:29004. doi: 10.1038/srep29004.

Birrell P, Blake J, van Leeuwen E, PHE Joint Modelling Cell, Gent N, De Angelis D. Real-time nowcasting and forecasting of COVID-19 dynamics in England: the first wave? *medRxiv* [preprint] 2020.08.24.20180737; doi: https://doi.org/10.1101/2020.08.24.20180737.

Bishop MT, Diack AB, Ritchie DL, Ironside JW, Will RG, Manson JC. Prion infectivity in the spleen of a PRNP heterozygous individual with subclinical variant Creutzfeldt-Jakob disease. *Brain* 2013;135(4):1139-1145. doi: 10.1093/brain/awt032.

Clarke P, Will RG, Ghani AC. Is there the potential for an epidemic of variant Creutzfeldt-Jakob disease via blood transfusion in the UK? *J R Soc Interface*. 2007;4(15):675-84. doi: 10.1098/rsif.2007.0216.

Collaborative Group on AIDS Incubation and HIV Survival including the CASCADE EU Concerted Action. Time from HIV-1 seroconversion to AIDS and death before widespread use of highly-active antiretroviral therapy: a collaborative re-analysis. *Lancet* 2000; 355: 1131-1137.

Collinge J, Sidle KC, Meads J, Ironside J, Hill AF. Molecular analysis of prion strain variation and the aetiology of 'new variant' CJD. *Nature*. 1996;383(6602):685-90. doi: 10.1038/383685a0.

Cooper JD, Bird SM. Predicting incidence of variant Creutzfeldt-Jakob disease from UK dietary exposure to bovine spongiform encephalopathy for the 1940 to 1969 and post-1969 birth cohorts. *Int J Epidemiol*. 2003;32(5):784-91. doi: 10.1093/ije/dyg248.

da Silva SG, Leon LA, Alves G, Brito SM, Sandes Vde S, Lima MM, Nogueira MC, Tavares Rde C, Dobbin J, Apa A, de Paula VS, Oliveira JM, Pinto MA, Ferreira Oda C Jr, Motta Ide J. A Rare Case of Transfusion Transmission of Hepatitis A Virus to Two Patients with Haematological Disease. *Transfus Med Hemother*. 2016;43(2):137-41. doi: 10.1159/000441910.

Detels R, Muñoz A, McFarlane G, Kingsley LA, Margolick JB, Giorgi J, Schrager LK, Phair JP for the Multicentre AIDS Cohort Study Investigators. Effectiveness of potent antiretroviral therapy on time to AIDS and death in men with known HIV infection duration. Multicenter AIDS Cohort Study Investigators. *JAMA 1*998;280(17):1497-503. doi: 10.1001/jama.280.17.1497.

Dunn JJ, Starke JR, Revell PA. Laboratory Diagnosis of Mycobacterium tuberculosis Infection and Disease in Children. *J Clin Microbiol*. 2016;54(6):1434-1441. doi: 10.1128/JCM.03043-15.

Editorial. Is it time to rethink UK restrictions on blood donation? *EClinicalMedicine*. 2019;15:1-2. doi: 10.1016/j.eclinm.2019.10.014.

Gill ON, Spencer Y, Richard-Loendt A, Kelly C, Dabaghian R, Boyes L, Linehan J, Simmons M, Webb P, Bellerby P, Andrews N, Hilton DA, Ironside JW, Beck J, Poulter M, Mead S, Brandner S. Prevalent abnormal prion protein in human appendixes after bovine spongiform encephalopathy epizootic: large scale survey. *BMJ* 2013;347:f5675. doi: 10.1136/bmj.f5675.

Gore SM. Ban on home HIV tests is unjustified. *BMJ*. 1992;304(6834):1118. doi: 10.1136/bmj.304.6834.1118-c.

Greenwood M. The epidemiology of influenza. *Br Med J*. 1918;2(3021):563-6. doi: 10.1136/bmj.2.3021.563.

Gudbjartsson DF, Norddahl GL, Melsted P, Gunnarsdottir K, Holm H, Eythorsson E, Arnthorsson AO, Helgason D, Bjarnadottir K, Ingvarsson RF, Thorsteinsdottir B, Kristjansdottir S, Birgisdottir K, Kristinsdottir AM, Sigurdsson MI, Arnadottir GA, Ivarsdottir EV, Andresdottir M, Jonsson F, Agustsdottir AB, Berglund J, Eiriksdottir B, Fridriksdottir R, Gardarsdottir EE, Gottfredsson M, Gretarsdottir OS, Gudmundsdottir S, Gudmundsson KR, Gunnarsdottir TR, Gylfason A, Helgason A, Jensson BO, Jonasdottir A, Jonsson H, Kristjansson T, Kristinsson KG, Magnusdottir DN, Magnusson OT, Olafsdottir LB, Rognvaldsson S, le Roux L, Sigmundsdottir G, Sigurdsson A, Sveinbjornsson G, Sveinsdottir KE, Sveinsdottir M, Thorarensen EA, Thorbjornsson B, Thordardottir M, Saemundsdottir J, Kristjansson SH, Josefsdottir KS, Masson G, Georgsson G, Kristjansson M, Moller A, Palsson R, Gudnason T, Thorsteinsdottir U, Jonsdottir I, Sulem P, Stefansson K. Humoral Immune Response to SARS-CoV-2 in Iceland. *N Engl J Med*. 2020;383(18):1724-1734. doi: 10.1056/NEJMoa2026116.

Hill AF, Desbruslais M, Joiner S, Sidle KC, Gowland I, Collinge J, Doey LJ, Lantos P. The same prion strain causes vCJD and BSE. *Nature*. 1997;389(6650):448-50. doi: 10.1038/38925.

Hilton DA, Fathers E, Edwards P, Ironside JW, Zajicek J. Prion immunoreactivity in appendix before clinical onset of variant Creutzfeldt-Jakob disease. *Lancet*.1998;352(9129):703-4. doi: 10.1016/S0140-6736(98)24035-9.

Hilton DA, Ghani AC, Conyers L, Edwards P, McCardle L, Ritchie D, Penney M, Hegazy D, Ironside JW. Prevalence of lymphoreticular prion protein accumulation in UK tissue samples. *J Pathol*. 2004;203(3):733-9. doi: 10.1002/path.1580.

Hutchinson SJ, Bird SM, Goldberg DJ. Modeling the current and future disease burden of hepatitis C among injection drug users in Scotland. *Hepatology*. 2005;42(3):711-23. doi: 10.1002/hep.20836.

Hutchinson SJ, Valerio H, McDonald SA, Yeung A, Pollock K, Smith S, Barclay S, Dillon JF, Fox R, Bramley P, Fraser A, Kennedy N, Gunson RN, Templeton K, Innes H, McLeod A, Weir A, Hayes PC, Goldberg D. Population impact of direct-acting antiviral treatment on new presentations of hepatitis C-related decompensated cirrhosis: a national record-linkage study. Gut. 2020;69(12):2223-2231. doi: 10.1136/gutjnl-2019-320007.

Kay AW, González Fernández L, Takwoingi Y, Eisenhut M, Detjen AK, Steingart KR, Mandalakas AM. Xpert MTB/RIF and Xpert MTB/RIF Ultra assays for active tuberculosis and rifampicin resistance in children. Cochrane Database Syst Rev. 2020;8(8):CD013359. doi: 10.1002/14651858.CD013359.pub2.

Kilbourne ED. Influenza pandemics of the 20th century. *Emerg Infect Dis*. 2006;12(1):9-14. doi: 10.3201/eid1201.051254.

Llewelyn CA, Hewitt PE, Knight RS, Amar K, Cousens S, Mackenzie J, Will RG. Possible transmission of variant Creutzfeldt-Jakob disease by blood transfusion. *Lancet*. 2004;363(9407):417-21. doi: 10.1016/S0140-6736(04)15486-X.

Miller E, Hoschler K, Hardelid P, Stanford E, Andrews N, Zambon M. Incidence of 2009 pandemic influenza A H1N1 infection in England: a cross-sectional serological study. *Lancet*. 2010.27;375(9720):1100-8. doi: 10.1016/S0140-6736(09)62126-7.

Murphy SC, Shott JP, Parikh S, Etter P, Prescott WR, Stewart VA. Malaria diagnostics in clinical trials. *Am J Trop Med Hyg*. 2013;89(5):824-39. doi: 10.4269/ajtmh.12-0675.

Ngasala B, Bushukatale S. Evaluation of malaria microscopy diagnostic performance at private health facilities in Tanzania. *Malar J*. 2019 Nov 26;18(1):375. doi: 10.1186/s12936-019-2998-1.

Nicol MP, Zar HJ. New specimens and laboratory diagnostics for childhood pulmonary TB: progress and prospects. *Paediatr Respir Rev*. 2011 Mar;12(1):16-21. doi: 10.1016/j.prrv.2010.09.008.

Peden AH, Head MW, Ritchie DL, Bell JE, Ironside JW. Preclinical vCJD after blood transfusion in a PRNP codon 129 heterozygous patient. *Lancet*. 2004;364(9433):527-9. doi: 10.1016/S0140-6736(04)16811-6.

Public Health England. Guidance HIV: self-testing, 1 April 2014. Available at: https://www.gov.uk/government/publications/hiv-self-testing [accessed 21 May 2021].

Sekine T, Perez-Potti A, Rivera-Ballesteros O, Strålin K, Gorin JB, Olsson A, Llewellyn-Lacey S, Kamal H, Bogdanovic G, Muschiol S, Wullimann DJ, Kammann T, Emgård J, Parrot T, Folkesson E; Karolinska COVID-19 Study Group, Rooyackers O, Eriksson LI, Henter JI, Sönnerborg A, Allander T, Albert J, Nielsen M, Klingström J, Gredmark-Russ S, Björkström NK, Sandberg JK, Price DA, Ljunggren HG, Aleman S, Buggert M. Robust T Cell Immunity in Convalescent Individuals with Asymptomatic or Mild COVID-19. *Cell*. 2020;183(1):158-168.e14. doi: 10.1016/j.cell.2020.08.017.

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

Spearman CW, Dusheiko GM, Hellard M, Sonderup M. Hepatitis C. *Lancet*. 2019.19;394(10207):1451-1466. doi: 10.1016/S0140-6736(19)32320-7.

Struble K, Chan-Tack K, Qi K, Valappil T, Connelly S, Mishra P, Price D, Murray J, Birnkrant D. Sustained virological response rates with direct-acting antivirals in black subjects with HCV genotype 1 infection: systematic analysis of clinical trials. *J Virus Erad*. 2019;5(3):138-144.

To KK, Hung IF, Ip JD, Chu AW, Chan WM, Tam AR, Fong CH, Yuan S, Tsoi HW, Ng AC, Lee LL, Wan P, Tso E, To WK, Tsang D, Chan KH, Huang KH, Kok KH, Cheng VC, Yuen KY. Coronavirus Disease 2019 (COVID-19) Re-infection by a Phylogenetically Distinct Severe Acute Respiratory Syndrome Coronavirus 2 Strain Confirmed by Whole Genome Sequencing. *Clinical Infectious Diseases* 2020;ciaa1275. doi: 10.1093/cid/ciaa1275.

Van Damme P. Long-term Protection After Hepatitis B Vaccine. J Infect Dis. 2016 Jul 1;214(1):1-3. doi: 10.1093/infdis/jiv750.

Viviani S, Carrieri P, Bah E, Hall AJ, Kirk GD, Mendy M, Montesano R, Plymoth A, Sam O, Van der Sande M, Whittle H, Hainaut P; Gambia Hepatitis Intervention Study. 20 years into the Gambia Hepatitis Intervention Study: assessment of initial hypotheses and prospects for evaluation of protective effectiveness against liver cancer. *Cancer Epidemiol Biomarkers Prev*. 2008;17(11):3216-23. doi: 10.1158/1055-9965.EPI-08-0303.

Weiss GE, Traore B, Kayentao K, Ongoiba A, Doumbo S, Doumtabe D, Kone Y, Dia S, Guindo A, Traore A, Huang CY, Miura K, Mircetic M, Li S, Baughman A, Narum DL, Miller LH, Doumbo OK, Pierce SK, Crompton PD. The Plasmodium falciparum-specific human memory B cell compartment expands gradually with repeated malaria infections. *PLoS Pathog*. 2010;6(5):e1000912. doi: 10.1371/journal.ppat.1000912.

Will RG, Ironside JW, Zeidler M, Cousens SN, Estibeiro K, Alperovitch A, Poser S, Pocchiari M, Hofman A, Smith PG. A new variant of Creutzfeldt-Jakob disease in the UK. *Lancet*. 1996;347(9006):921-5. doi: 10.1016/s0140-6736(96)91412-9.

World Health Organization. 2009. *Methods for field trials of malaria rapid diagnostic tests*. December 2009. Available at: https://www.who.int/malaria/publications/atoz/9789290614166_field_trials/en/. [Accessed 21 May 2021].

World Health Organization. 2015. *Guidelines for the treatment of malaria. Third edition*. April 2015. Available at: https://apps.who.int/iris/handle/10665/162441 [Accessed 21 May 2021].

World Health Organization. 2016. *Malaria Microscopy Quality Assurance Manual – Version 2*. 2016. Available at: https://www.who.int/malaria/publications/atoz/9789241549394/en/. [Accessed 21 May 2021].

World Health Organization. 2017. *Protecting malaria high-risk groups (online)*. Available at: https://www.who.int/activities/protecting-malaria-high-risk-groups [Accessed 21 May 2021].

World Health Organization 2019(A). *Global tuberculosis report 2019*. Available at: https://www.who.int/tb/publications/global_report/en/. [Accessed 21 May 2021].

World Health Organization 2019(B). *Malaria report 2019*. Available at: https://www.who.int/publications/i/item/9789241565721. [Accessed 21 May 2021].

Wroe SJ, Pal S, Siddique D, Hyare H, Macfarlane R, Joiner S, Linehan JM, Brandner S, Wadsworth JD, Hewitt P, Collinge J. Clinical presentation and pre-mortem diagnosis of variant Creutzfeldt-Jakob disease associated with blood transfusion: a case report. *Lancet*. 2006;368(9552):2061-7. doi: 10.1016/S0140-6736(06)69835-8.

Young BE, Fong SW, Chan YH, Mak TM, Ang LW, Anderson DE, Lee CY, Amrun SN, Lee B, Goh YS, Su YCF, Wei WE, Kalimuddin S, Chai LYA, Pada S, Tan SY, Sun L, Parthasarathy P, Chen YYC, Barkham T, Lin RTP, Maurer-Stroh S, Leo YS, Wang LF, Renia L, Lee VJ, Smith GJD, Lye DC, Ng LFP. Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study. *Lancet*. 2020;396(10251):603-611. doi: 10.1016/S0140-6736(20)31757-8.

From past to present...
The image of the wheatsheaf first appeared in our original seal. Being the end product of the harvesting and bundling of wheat, it was a pictorial way of expressing the gathering and analysis of data: the foundations of statistical work.
It also implied that statistical practice comprises more than the collection of data: it consists of active interpretation and application as well (threshed for others, if the rural analogy is sustained). Rigorous data gathering is still at the heart of modern statistics, but as statisticians we also interpret, explain and present the data we collect.