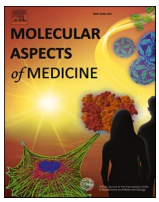
Molecular Aspects of Medicine 97 (2024) 101275

Contents lists available at ScienceDirect

Molecular Aspects of Medicine

journal homepage: www.elsevier.com/locate/mam



Review

Ensuring accuracy in the development and application of nucleic acid amplification tests (NAATs) for infectious disease

Jim F. Huggett [a,b,*], Denise M. O'Sullivan [a,b], Simon Cowen [a], Megan H. Cleveland [c], Kerrie Davies [d,e,f], Kathryn Harris [g], Jacob Moran-Gilad [h], Amanda Winter [i], Julian Braybrook [a], Michael Messenger [j,k,l]

[a] National Measurement Laboratory (NML), LGC, Queens Road, Teddington, TW11 0LY, Middlesex, UK
[b] School of Biosciences & Medicine, Faculty of Health & Medical Science, University of Surrey, Guildford, UK
[c] Material Measurement Laboratory, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899, USA
[d] Healthcare Associated Infections Research Group, Leeds Teaching Hospitals NHS Trust/University of Leeds, UK
[e] NIHR Leeds MedTech In Vitro Diagnostic Cooperative, University of Leeds, UK
[f] NIHR Leeds Biomedical Research Centre, Leeds Teaching Hospitals and University of Leeds, UK
[g] Department of Virology, NHS East and South East London Pathology Partnership, Royal London Hospital, Barts Health NHS Trust, London, UK
[h] Department of Health Policy and Management, School of Public Health, Faculty of Health Sciences, Ben Gurion University of the Negev, Beer Sheva, Israel
[i] The Newcastle Upon Tyne Hospitals NHS Foundation Trust, North East Innovation Laboratory, The Biosphere, Drayman's Way, Newcastle Upon Tyne, NE4 5BX, UK
[j] FIND, Campus Biotech, Chemin des Mines 9, 1202 Geneva, Switzerland
[k] School of Medicine and Health, University of Leeds Worsley Building, University of Leeds, Woodhouse, Leeds LS2 9JT, UK
[l] British In Vitro Diagnostic Association (BIVDA), 299 Oxford St, London, W1C 2DZ, UK

ARTICLE INFO

Keywords:
NAAT
Nucleic acid amplification test
PCR
Target product profile
TPP
Diagnosis
In vitro diagnostics
IVD
Molecular diagnosis
MDx
Infectious disease
Pathogen

ABSTRACT

Diagnostic tests were heralded as crucial during the Coronavirus disease (COVID-19) pandemic with most of the key methods using bioanalytical approaches that detected larger molecules (RNA, protein antigens or antibodies) rather than conventional clinical biochemical techniques. Nucleic Acid Amplification Tests (NAATs), like the Polymerase Chain Reaction (PCR), and other molecular methods, like sequencing (that often work in combination with NAATs), were essential to the diagnosis and management during COVID-19. This was exemplified both early in the pandemic but also later on, following the emergence of new genetic SARS-CoV-2 variants.

The 100 day mission to respond to future pandemic threats highlights the need for effective diagnostics, therapeutics and vaccines. Of the three, diagnostics represents the first opportunity to manage infectious diseases while also being the most poorly supported in terms of the infrastructure needed to demonstrate effectiveness. Where performance targets exist, they are not well served by consensus on how to demonstrate they are being met; this includes analytical factors such as limit of detection (LOD) false positive results as well as how to approach clinical evaluation. The selection of gold standards or use of epidemiological factors such as predictive value, reference ranges or clinical thresholds are seldom correctly considered.

The attention placed on molecular diagnostic tests during COVID-19 illustrates important considerations and assumptions on the use of these methods for infectious disease diagnosis and beyond. In this manuscript, we discuss state-of-the-art approaches to diagnostic evaluation and explore how they may be better tailored to diagnostic techniques like NAATs to maximise the impact of these highly versatile bioanalytical tools, both generally and during future outbreaks.

1. Introduction

Nucleic acid amplification tests (NAATs) and advanced sequencing have been increasingly applied to support clinical diagnosis and management of communicable and non-communicable diseases over the last decade. Despite this, the Coronavirus disease (COVID-19) pandemic-associated discussions around bioanalytical diagnostic techniques were unprecedented, with widespread reporting of these methods becoming

* Corresponding author. National Measurement Laboratory (NML), LGC, Queens Road, Teddington, TW11 0LY, Middlesex, UK.
E-mail address: Jim.Huggett@lgcgroup.com (J.F. Huggett).

https://doi.org/10.1016/j.mam.2024.101275
Received 15 November 2023; Received in revised form 29 March 2024; Accepted 22 April 2024
Available online 20 May 2024

everyday topics discussed by the public and professionals alike, on a global scale. Such molecular testing is also under increased scrutiny from guidance infrastructure and regulators as it becomes more established in clinical diagnostics fields ranging from precision medicine to foetal testing to infectious disease.

NAATs and other molecular techniques differ from other *in vitro* diagnostics (IVDs) because the polymer analytes in question (whether RNA, DNA or their modifications) are variations on the same theme. Consequently, many of the challenges faced when developing a NAAT to detect a genetic predictor to a cancer treatment are shared with those associated with a diagnostic test targeting a pathogen's genomic material to diagnose an infectious disease like tuberculosis. Yet there are also unique considerations for nucleic acid analytes when compared to other IVDs; COVID-19 arguably illustrated a wide regional discrepancy in understanding of the diagnostic capabilities afforded by NAATs.

While PCR is by far the most common NAAT used in research or diagnostic testing, other formats also exist, most of which are isothermal (i.e. unlike PCR, do not require thermocycling). Alternative methodologies may provide advantages such as speed, simplicity, near-patient application, cost reduction and affordability, especially in the face of PCR reagent scarcity. Alternative NAATs cannot easily outcompete a well-designed PCR protocol when considering limit of detection (LOD) (Huggett et al., 2021); this may prove a limitation depending on the target analyte's concentration (Fig. 1) and the resulting impact on disease diagnosis. Essentially, it is how an individual method has been designed and validated from the outset (rather than the NAAT format itself) in the context of the disease in question that will dictate its use as an IVD and whether alternative NAAT chemistries can provide unique

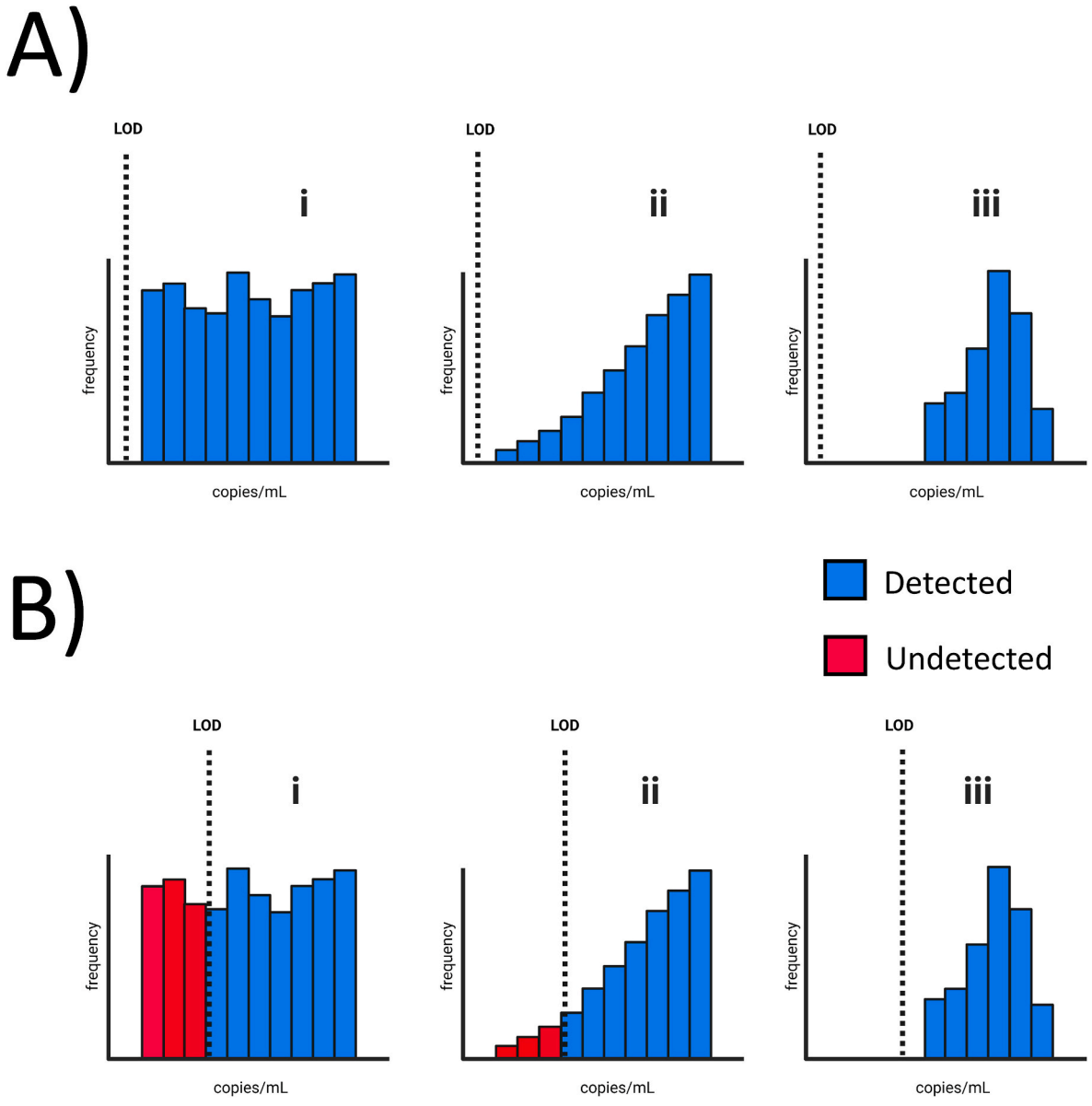


Fig. 1. Theoretical distribution of analytes and how different LODs will impact on sensitivity. (i) Shows a uniform distribution with even representation across the range (ii) left skewed distribution result with fewer low concentration specimens and (iii) normal distribution with results that are all at higher concentrations. When using a method with low LOD (A) the quantitative distribution has no effect on sensitivity as all samples are scored as detected. However, when using a method with higher LOD (B), the distribution of the analyte impacts on the proportion of samples scored detected (blue bars) or undetected (red bars). N.B. To aid in this discussion quantiles below the LOD are assigned as undetected, however in reality detection is possible below the LOD at a reduced frequency to that set for the LOD. How detected and undetected are treated may vary with specific clinical context, but this figure illustrates how knowledge of the quantitative distribution could guide the use of diagnostic tests with different LOD. SARS-CoV-2 RNA distribution varied between i and ii during the COVID-19 pandemic (Hay et al., 2021). Created with BioRender.com.

solutions in a world dominated by PCR.

As a diagnostic analyte, nucleic acid target sequences can vary in concentration in a biological specimen over a considerable dynamic range. For example, viral nucleic acids quantities can span over eight orders of magnitude from specimen to specimen (Hijano et al., 2019). PCR as a diagnostic method is also unique in that it can detect across most of this vast dynamic range, with correctly optimised PCR reactions able to detect in excess of 100 million DNA copies per reaction down to trace levels of target. This can be both a benefit and a hindrance depending on the diagnostic context.

Current evaluation of diagnostic performance using conventional methods for determination of test accuracy typically employs specificity, sensitivity and predictive values which are often poorly approached and misunderstood. Such approaches are dependent on a clinical reference standard definition of the target condition or disease state and an appropriately designed validation plan (Royal Statistical Society, 2021). However, the clinical reference standard often merges multiple analytical and clinical measures to aid classification which are often imperfect. This is especially the case in the early stages of understanding a disease, such as when a novel pathogen emerges. Arguably a more dynamic approach to setting and evaluating performance targets is needed, especially during an emergency context, to maximise the impact of NAATs and other molecular methods.

Evaluation of the analytical performance which influences sensitivity (dynamic range, LOD, inclusivity) and specificity (exclusivity, cross-reactivity) of the molecular test along with clinical reference ranges can aid understanding and optimisation of potential clinical performance prior to clinical evaluation of a given IVD. However, while requirements for these analytical criteria are generally stipulated in target product profile (TPPs) (e.g. for SARS-CoV-2 (WHO, 2020), mpox (WHO, 2023), etc.) recommendations on how to set and evaluate tests against such criteria remain unclear (Sandberg et al., 2015). The clinical reference ranges of pathogen nucleic acids are also seldom investigated or defined to aid in understanding the quantities of analyte available for detection by a given analytical solution. This means the required dynamic range or LOD for a given clinical sensitivity is unknown, requiring guesswork when setting targets. This is further complicated when an assay is designed for the purpose of quantification, rather than mere detection of the said target.

Discussions around molecular diagnostic performance are exacerbated by the fact that, unlike more established areas of clinical chemistry, nucleic acid analysis is less well supported in terms of material and methodological standards, to deliver traceability to underpin routine detection and quantification. Diagnostic detection of DNA and RNA sequences is increasingly common and relative quantification to an internal genetic reference to define a clinical threshold is also expanding. However, the widespread applied quantification, where a clinical decision is made based on the absolute abundance of that nucleic acid (per unit volume) is quite rare, limited to a handful of viral targets including human immunodeficiency virus (HIV), hepatitis C virus (HCV), hepatitis B virus (HBV) and human cytomegalovirus (hCMV). This may, in part, be due to challenges associated with the standardisation that is required to underpin such measurements. There is also a common misconception that if a method is not 'quantitative,' and just detects the presence of a specific sequence, it is simple to standardise; this may or may not be the case depending on the clinical situation being investigated and abundance of the nucleic acid of interest (Fig. 1).

This manuscript explores the preceding points and uses the experience from COVID-19 to discuss how PCR and other molecular methods could be improved. We discuss how analytical considerations could be better applied to aid decisions in the deployment of bioanalytical diagnostic methods in the future, as part of an emergency outbreak response, but also in the wider use to molecular methods to meet the increased requirement for regulation. We provide a series of recommendations which we present in Table 1.

Table 1
Summary recommendations for NAAT diagnostic performance assessment when used to detect the presence/absence of pathogen nucleic acids.

Topic	Challenge	Recommendation
Link between LOD and clinical sensitivity	LOD can be directly linked to clinical sensitivity (Fig. 1). Methods with different LODs will categorise patients as positive or negative if analyte quantities fall between the difference in LOD.	Knowledge of the analyte clinical reference range will aid in determining the required LOD for a given test. The analyte quantitative distribution can be evaluated using reference methods such as calibrated quantitative real time PCR (qPCR) or digital PCR (dPCR). In the absence of more accurate methods data from uncalibrated NAAT outputs (e.g. quantification cycle (C_q) or cycle threshold (C_t)) may be used to provide a crude estimation of distribution using populations of positive results.
Gold standards for clinical evaluation	Clinical evaluation studies often select specimens as positive and negative using gold standards methods. This may also be done without knowledge of analyte quantities within the respective specimens.	Gold standards methods (usually alternative diagnostic tests) can provide a useful comparator, but the analyte concentration within specimens used for evaluation of sensitivity should be considered as this may impact on the reproducibility of the findings. Gold standards methods are also of limited value if the new method outperforms the comparator approach. Accurate reference methods that may not be intended for routine diagnosis can also be considered to characterise the specimens (pathogen nucleic acid sequence and quantity) prior to clinical evaluation.
Impact of prevalence on diagnostic performance	Predictive values are often calculated using the results from clinical evaluation studies (sensitivity and specificity)	Predictive value should be calculated based on the prevalence of the condition in those being tested and not from the clinical evaluation study which will have an artificial prevalence deliberately powered to allow performance assessment. Positive predictive value is usually the metric that is of concern as prevalence is usually low meaning false positive results will have a higher impact at the population level.
Impact of incidence on diagnostic performance	Changes in incidence can influence the abundance of nucleic acids within a given population of clinical specimens resulting in a sensitivity shift.	The sensitivity shift could be used to guide the selection of different test formats (with differing analytical performance criteria, such as LOD) to different stages of an outbreak.
False positive results	NAATs are especially susceptible to false positive results.	Consider the fact that NAAT false positive results are usually caused by contamination from NAAT amplicon, other specimens or synthetic template as the first port of call when trouble shooting.
Setting of quantitative thresholds	Pathogen nucleic acid quantities may correlate with disease severity,	If quantitative nucleic acid sequence thresholding can be demonstrated to be clinically

(*continued on next page*)

Table 1 (*continued*)

Topic	Challenge	Recommendation
	infectiousness or predict outcome. This is not a given however and cannot be assumed. Furthermore, thresholds should not simply be applied to any NAAT result. Nucleic acid thresholding is a quantitative measurement and the methods used for this purpose must be optimised for that purpose with an appropriate calibration setup. Multiplex NAAT tests that have not been optimised for quantification may not show good linearity.	significant for patient stratification, then wider application will required a calibration strategy. The setting of an un-calibrated NAAT result as a threshold, such as above or below C_q (C_t) 25, across different laboratories should be avoided.

2. Diagnostic, therapeutics, and vaccines (DTVs)

The "100 days mission to respond to future pandemic threats" introduced by Sir Patrick Vallance and Melinda Gates was a report to the G7 (first published on 12th June 2021 (G7, 2021)) and is an example of a number of international and national initiatives for better pandemic preparedness. It is an ambitious call to respond to a potential epidemic or pandemic threat at pace to prevent wider spread by deploying a range of medical and non-medical interventions. On the topic of medical interventions, the mission calls for "safe, effective and affordable" diagnostic tests, therapeutics and vaccines (together termed DTVs) in the first 100 days of a pandemic threat being identified: defined by when World Health Organization (WHO) declares a public health emergency of international concern (PHEIC). Of the three, therapeutics and vaccines get the most attention and funding, as they comprise the interventions that can mitigate the effects of the disease in question, or potentially stop the pandemic in its tracks. The implementation of therapeutics and vaccines is also comprehensively regulated to maximise their potential while reducing potential harm from unintended side effects and potential for serious acute adverse events in patients.

Diagnostics have an interesting place within the DTVs as being both seen as "crucial" (Peeling et al., 2022) while also far less well supported in terms of funding, regulation, or analytical consensus. It is important to note that while the speed of development and efficacy of therapeutics and vaccines for SARS-CoV-2 was extraordinary, it would be unwise to assume this could be repeated for every potential pandemic pathogen. Indeed, we have been unable to develop a vaccine for HIV since it's discovery in 1983, and many other pathogen vaccines are similarly without success. It is easy to imagine a situation where correctly identifying disease cases using IVDs and implementing non-pharmaceutical interventions (NPIs) is the only DTV solution to a future outbreak, at least in its initial stages. The very early deployment of diagnostics during COVID-19, supported a similar role during the first year of the pandemic.

Many of the pandemic response documents, including the 100-day mission, use terms like 'standardised' or 'approved' when describing diagnostic assays. However, what is not clear is what these terms refer to; such as how they need to be 'approved,' or how 'standarised' should be demonstrated and whether this is considering material, methodological or documentary standards. This is not helped by the fact that, unlike for therapeutics and vaccines, there is limited consensus in how diagnostic performance evaluations should be conducted in an emergency context and what is the sufficient level of evidence that supports the roll out of diagnostics.

The above point is arguably exacerbated by the fact that bioanalytical techniques like NAATs (as well as antigen and serological tests) are often more difficult to standardise. Routes to improve this are discussed below, but first we will explore the current approaches to setting diagnostic targets when using molecular methods like NAATs.

3. Diagnostic specification documents. The example of the target product profile (TPP)

Early in an outbreak many nations rely upon emergency use authorization (EUA) powers to facilitate market access for urgently needed medical devices, including IVDs. Whilst the specific legal frameworks and processes for EUA's differ by jurisdiction, they generally share similarities in providing for regulatory flexibility to supply non-compliant medical devices on humanitarian grounds. EUAs are usually granted on the basis of a risk: benefit analysis by the regulator and typically impose conditions upon the manufacturer to achieve compliance within a defined period of time. Early in the COVID-19 pandemic, regulatory requirements for EUAs were more fluid and based upon the best available information available at the time, as a result the minimum and optimal requirements for safety and performance were not always clear to manufacturers. Over time and in response to the evolving evidence and policy landscape, regulators in partnership with wider stakeholders, strived to clarify and better articulate the necessary requirements for safe and effective medical products. TPPs (e.g. (MHRA, 2023; WHO)) represent examples of such diagnostic specification documents that are used to articulate the requirements of an IVD. Other examples include European Union Common Specifications (European Union, 2022), Harmonised Standards (e.g. ISO 15197:2013; CLSI POCT05), FDA Templates for Developers (FDA, 2021), WHO Technical Specification Series (WHOb) and Preferred Product Characteristics (PPC).

TPPs are described by the WHO as documents that "outline the desired 'profile' or characteristics of a target product that is aimed at a particular disease or diseases." These documents, used to guide the development of medical technologies, stipulate the minimal and preferred requirements for key test characteristics, based upon expert consensus of the available evidence at a given moment in time. A TPP for a diagnostic test will typically define criteria (along with targets) such as intended use, population, specimen type, time to result, etc. The key criteria specified that are linked to the desired diagnostic accuracy of a test are the analytical and clinical performance which are often described with 'preferred' or 'minimal' targets (Table 2).

Analytical performance can be defined broadly as the criteria that describe how well the method detects the analyte (further narrowed down to measurand) in question. This is completely independent of how appropriate or accurate that result may be for the intended clinical use. For molecular tests like NAATs analytical performance includes considerations of:

- the dynamic range over which the range of quantities of analyte that can be detected and/or quantified. Crucially this is influenced by more than just the final analytical step; with heavy influence from the pre-examination steps including sampling, storage and transport and extraction (Mercer et al., 2022).
- LOD and in some cases limit of quantification (LOQ) essentially describe the lower concentration of the dynamic range and define when the concentration becomes too low to be accurately detected (or quantified). As LOD (and LOQ) are linked to dynamic range it is also impacted by the pre-examination and final analytical step. However, for many NAATs LOD can also be impacted by an additional factor: physics. Given that NAATs like PCR can detect single molecules their presence can be the factor that determines LOD. This is exemplified by dPCR (dMIQE and Huggett, 2020). LOD is arguably one of the most important analytical parameters when considering sensitivity as, depending on the quantitative distribution of the analyte within the clinical specimen and clinical meaning of the result at a given analyte concentration, it can be intrinsically linked to and govern clinical sensitivity (see Fig. 1 and Section 5.3).

Table 2
Example of Performance and Regulatory criteria targets for tests used for mpox diagnosis within health care settings and laboratories https://www.who.int/publications/i/item/9789240076464.

Characteristic	Minimal	Preferred	Comment(s)
Analytical performance	1. Inclusivity: Able to detect clades I, IIa and IIb. 2. Limit of detection (LOD): determined using control material of defined quantity, equivalent to at least 1000 genomic copies per ml of specimen. 3. Analytical specificity: -assay performance should not be impacted by common interfering substances - assay should not cross-react with other common human pathogens, especially those causing similar signs and symptoms as MPXV (e.g., VZV, HSV). -MPXV specific target(s), at least one per assay, should not cross-react with other closely-related human pathogens, especially those causing similar signs and symptoms as MPXV (e.g., VZV, HSV). - MPXV specific target(s), at least one per assay, should not cross-react with other closely-related human OPXV, e.g., Vaccinia virus (VACV), Cowpox virus (CPXV).		LOD is a quantitative measurement determined using control material of defined quantity. Example control materials for NAAT include synthetically derived nucleic acids in buffered solution, MPXV DNA and inactivated whole virus. Consequently, the type of material used for LOD assessment and the method for value assignment of that material's quantity should be included in any report on method LOD.
Clinical Sensitivity	≥95% when using lesion material compared to a reference molecular method.	≥97% when using lesion material compared to a reference molecular method.	Performance targets should be met for lesion material and ideally should be demonstrated using prospective or retrospective (remnant) natural clinical samples. Samples should cover a range of clinically relevant viral loads, e.g., C_t equivalent 15–38 as per the reference method.
Clinical Specificity	≥97% when using lesion material compared to a reference molecular method.	≥99% when using lesion material compared to a reference molecular method.	
Invalid/error rate	≤5%[a]		
Manufacturing/ Regulatory approvals	ISO 13485:2016 compliant		ISO 13485:2016 compliant AND 1) WHO prequalification or WHO emergency use listing (as available) AND/OR 2) Authorization by a founding member of the Global Harmonization Task Force (Australia, Canada, European Union, Japan, United States of America)

[a] This invalid/error rate target is set to ensure a comprehensive set of controls are incorporated.

In addition to criteria associated with detection of intended analyte the analytical performance may also consider different strains, clades, genetic variants etc to be detected and to ensure the method in question is analytically specific with regards to cross-reactivity (the level to which it may detect other related or unrelated microorganisms that may be present) as well as the potential impact of potential sources of interference (influence of other substances found in the sample matrix on the test result).

Clinical performance criteria consider what the results of a method (based on its given analytical performance) mean from an operational perspective (clinical or epidemiological). Clinical performance of a given diagnostic test is typically articulated using sensitivity (percentage of patients with the condition correctly identified by the test as having the condition) and specificity (percentage of patients without the condition correctly identified by the test as not having the condition). A method that is analytically accurate may be clinically inaccurate because of an uncertain clinical correlation or in the case of SARS-CoV-2 the positive signal can still detect RNA when the patient has clinically recovered from the infection (Wu et al., 2021). While this point is logical when discussing diagnostic performance in theory, understanding its impact in practice is severely hampered by the lack of guidance on how to meet analytical performance criteria, combined with the challenges associated with defining positive and negative patients to support the definition of clinical targets (a fact that is exacerbated during a newly introduced infectious disease) or in empirically ascertaining the clinical significance of test results.

The above also serves as a reminder that a diagnostic test is typically used by a physician or other healthcare practitioner in combination with other supporting information (e.g. time from first symptoms or case-contact) to manage the patient i.e. it is the physician, not the test, that diagnoses the patient. This arguably allows for more variability in test performance as the physician is often armed with other sources of information on which to guide their decisions; so in the example of a positive SARS-CoV-2 result following recovery, the physician could use the patient's history and examination or timeline of infection to guide decision. However, the diagnostic envelope that is being pushed by NAATs challenges this paradigm, as molecular methods can detect pathogens, drug resistance or genotype cancers where there may be little or no validated actionable information on which to guide a clinical decision. Furthermore, when a test is utilised for epidemiological purposes, rather than clinical management (e.g. guiding NPIs or public health policy), supporting information may not be available or relevant, rendering test results as a key decision-support measure. Under these circumstances the need for robust tests performing within clear analytical criteria, along with the infrastructure that ensure they are performing within those defined parameters, will be of increasing importance.

4. NAAT analytical evaluation

4.1. Specificity (cross-reactivity)

Cross-reactivity is an important consideration because if the NAAT assay detects the wrong target it will incorrectly identify a patient as positive for a different pathogen or non-pathogenic microbe. This is a rather obvious consideration when developing a diagnostic test, with the ISO technical specification ISO/TS 5798:2022 on 'Requirements and recommendations for detection of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) by nucleic acid amplification methods' suggesting possible pathogens for testing cross-reactivity ranging from influenza to *Mycobacterium tuberculosis to Cryptococcus.* The WHO TPP for mpox stipulates the assay should not cross-react with other common human pathogens, including varicella zoster virus (VZV), herpes simplex virus (HSV) (WHO, 2023). These documents, and others, imply manufacturers need to source a range of alternative pathogen nucleic acid to experimentally check for cross-reactivity.

There are situations where closely related species/strains may be equally detected by an individual assay resulting in non, or less, clinically relevant positive results. In these situations more than one assay may be required to exclude non-relevant results such as with pertussis molecular testing or when using NAAT methods to predict resistance for Methicillin-resistant *Staphylococcus aureus* (MRSA). However, in many cases it is impossible to design an 'all inclusive' NAAT assay to detect across distant pathogens that may create similar symptoms as clinically implicated pathogens may not have common ancestors. For example, while there are examples of assays to SARS-CoV-2 that will also detect SARS-CoV-1, influenza has no such homology so a common assay cannot

be designed between SARS and influenza (or for that matter mpox and VZV): so it is highly unlikely that one will be accidently developed. This is not to state that bioinformatics alone would be sufficient for evaluating assay specificity or that *in vitro* wet lab evaluation of specificity is not important (primarily to ensure the assay does not cross-react with human nucleic acids). Instead new assays could arguably depend more on *in silico* bioinformatics to consider the pathogen in question where cross-reactivity is more likely rather than blindly following a large-scale costly evaluation of multiple microbial extracts that we already know through genomic information cannot be detected by a specific NAAT.

4.2. False positive results

For the majority of infectious disease molecular diagnostic tests, the major source of false positives when conducting NAATs is contamination. This is because NAATs are capable of very sensitive detection, in some cases to the level of a single molecule (Yoo et al., 2016), and most of them achieve this by generating billions of copies of the very sequence of interest (termed the amplicon). Consequently, the products of previous tests constitute the major contamination threat to subsequent analyses. Sources of false positives can also occur when the oligonucleotide producers of the primers and probes also makes synthetic template molecules (Huggett et al., 2020). In addition to amplicon contamination, cross contamination from high titre specimens (especially viruses) to negative specimens, is also a risk during specimen handling. Routes to manage NAAT application to reduce contamination are outlined in ISO/DIS 17822:2020 (ISO, 2020). These include separate zones and creating uni-directional workflow within the laboratory for handling the different stages of the test, especially to prevent post experiment materials (amplicons) from contaminating subsequent experiments. Additional mitigation such as incorporating uracil bases and uracil DNA glycosylase have been used to reduce the impact of amplicon contamination. Negative controls are a fundamental consideration for monitoring the potential existence of contamination. Contamination is an important source of false positive results but represents a variable technical artefact that will differ between laboratories rather than a factor influencing specificity in the classical sense.

4.3. Dynamic range and limit of detection

The dynamic range of most NAATs when detecting presence/absence of a sequence is vast, capable of detecting nucleic acids across several orders of magnitude. Quantification across this range is possible with most real time PCR instruments (Bustin et al., 2009), although other formats (such as dPCR (dMIQE Group, 2020)) may require dilution due to the possibility of saturation. When developing a NAAT IVD, one of the major considerations for analytical requirements will arguably be the 'desirable' or 'minimal' required LOD. Improvements to LOD can be delivered through a variety of routes including assay design and optimisation, as well as choice of instrument, reagents and NAAT method. Finally, the type and volume/mass of specimen used, the sampling method as well as choice of extraction procedure can also have a major influence on the amount of nucleic acid available to the final experimental step (Fig. 2) and thus also influence LOD.

For the WHO TPPs for NAAT diagnosis of COVID-19 (WHO, 2020) and mpox ((WHO, 2023) and Table 2) the desired LOD in copies is currently 1000 copies/mL of specimen (typically a swab in appropriate buffer). When considering the detection of nucleic acids, the LOD describes the smallest quantity of DNA or RNA a method can reliably detect (within a given statistical criterion, usually 95%) and, importantly, this can be directly linked to the clinical sensitivity (depending on the quantitative distribution of the analyte in question within the population being tested for the target condition/disease). While LOD is a statistical term there are different ways to approach defining it and guidance on this when considering TPPs or wider methodological discussion can be explored further.

NAATs like PCR may be capable of (near-) single molecule detection, meaning the limit of detection may actually be determined by the probability of physical presence (or not) of the nucleic acid sequence within the test reaction. This is reflected in the criteria mentioned above

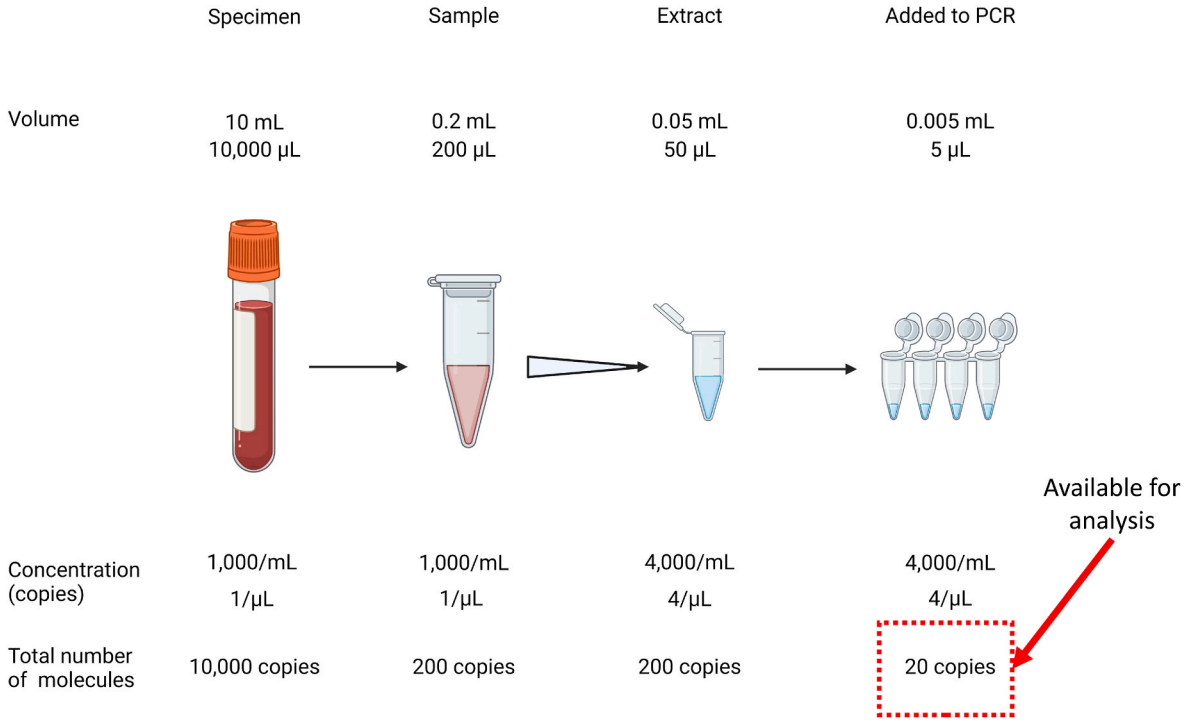


Fig. 2. Outline of how nucleic acid analyte quantity varies during typical NAAT workflow from original specimen to final experiment. Example is based on common LOD target of 1000 copies/mL of original specimen. This example assumes no loss (due to degradation or procedural inefficiencies) of analyte during the various steps from specimen taking to NAAT analysis and thus reflects highest possible amount available for analysis (red dotted box). Created with BioRender.com.

(and in Table 2) which stipulate 1000 copies/mL as minimal LOD. It is widely understood that this number is seen as a stringent bar for IVD manufacturers to meet in the absence of high-quality clinical evidence. However, when considering a typical protocol that may take 200 μL of specimen as starting material eluted into 50 μL of nucleic acid extract of which 5 μL will be analyzed, a 1000 copies/mL will correspond to, on average, 20 nucleic acid molecules in the experiment (Fig. 2).

When we consider the example in Fig. 2, assuming all preanalytical steps (specimen handling during analytical phase as well as its acquisition, transport and storage) and extraction are working perfectly, then 1000 copies/mL is a challenging target. This is because preanalytical steps will unlikely perform with 100% efficiency, further reducing the number of molecules available to the final NAAT reaction. Such a stringent LOD requirement could pose a barrier to developments that seek to improve throughput or point of care formats by reducing volumes, removing extraction (and thus specimen concentration) or switching to alternative technologies that may be less sensitive than PCR (Nixon et al., 2014).

Common examples where tests with reduced LODs were sought during COVID-19 included lateral flow antigen tests that had higher LODs than PCRs for inherent methodological reasons (antigen detection cannot detect as low a concentration of protein molecules as PCR can detect amplifiable DNA molecules) or due to procedural adaptations such as 'direct' PCR (Mercier et al., 1990) which sought to speed up the protocol by removing the extraction step. As the extraction also concentrates the nucleic acid (Fig. 2), 'direct' methods cannot benefit from the resultant improvement in LOD (Huggett et al., 2021); depending on analyte distribution this may or may not impact on test sensitivity (Fig. 1). Ultimately it is the biological dynamics of the analyte(s) in question, and how they relate to the disease in question, that dictates what method format (and minimal LOD) is needed for the most accurate diagnostic test; as outlined in section 5.3 below, characterising this relationship should be a priority when selecting targets for TPPs.

Furthermore, while TPPs (and similar documents) may request minimal or desirable targets to demonstrate that a test has met a given target (such as LOD) there are at least three additional challenges which need to be addressed to better aid the community in responding to such targets.

Firstly, paucity of guidance. While methods for calculating LOD are available (Armbruster and Pry, 2008) guidance within the TPP on what must be done to demonstrate a test has met a given target LOD is limited. At the extreme this could be split into:

A) statistical evidence that the test in question has an LOD significantly below a defined target copy number. Given many molecular tests are non-quantitative such an approach will require dilutions of accurately defined materials, considerable replication and, depending on confidence required, and distribution of the experimental results, an LOD that is several-fold below the target LOD such that the upper 95% boundary does not cross 1000 copies/mL, to assure it is met. This could make the target of 1000 copies/mL, or 20 per reaction, considerably more challenging.
B) statistical evidence (within defined experimental replicate criteria) that the test in question cannot be demonstrated to not have an LOD above the defined target copy number. This approach, which is used in analogous situation in the food testing sector (Burns et al., 2016), is simpler to meet both in terms of experimental design, resources and interpretation which in turn makes it easier to standardise and regulate.

Secondly, an absence of systems to support traceable measurement. Under LOD the WHO's mpox TPP states that "LOD is a quantitative measurement determined using control material of defined quantity" and that type of material used to determine LOD should "be included in any report on method LOD" (Table 2). This is an important inclusion as it places the onus on the IVD manufacturer to explain what materials were used to evaluate the LOD. This increases the importance of accurately defining this parameter beyond just a formality that is required to meet a specification, but still requires the manufacturer to determine how to meet this requirement which will likely lead to discrepancies.

Finally, infections reflect dynamic biological processes within tissue (s) and outside the infected cells. The nucleic acid quantities present may vary across the genome depending on a variety of factors leading to copy number variations. This is possibly most striking when measuring RNA for a virus like SARS-CoV-2 which uses RNA both for its genome and also for expression of additional protein coding RNAs termed sub genomic (sg)RNAs (Kim et al., 2020). This results in a 3-prime bias across the genome such that some specimens can have up to an order of magnitude higher RNA when comparing the 5-prime with the 3-prime sequences. As the NAAT cannot easily distinguish genomic RNA from sgRNA this raises the questions which sequence might the 1000 copies/mL guidance actually refer to and how the control materials that may be used to determine this LOD compare with real clinical specimens.

What is required to support the development of IVDs, both to meet regulatory analytical specifications but also for a more accurate definition of analytical performance, is accessibility to traceable reference materials (Page et al., 2020) which should be considered equally as part of a comprehensive diagnostic response to public health emergencies, alongside the development of new assays. Ideally these materials will have been characterised in terms of analyte quantity (and homogeneity), stability, identity and commutability. Wider attention is needed across the community spanning academic research, to IVD manufacturers, to regulators, on the systems to deliver traceability and underpin analytical accuracy in bioanalytical diagnostic development, implementation, and post-marketing surveillance. This is arguably needed if bioanalytical diagnostics are to take place alongside therapeutics and vaccines in pandemic response, or in other areas of medical diagnostic advancement such as in precision medicine or syndromic testing.

5. Clinical evaluation

5.1. Performance, sensitivity and specificity

Clinical performance defines how a test of given analytical performance categorises those being tested as having (or developing) the condition or not. For diagnostic tests using NAATs, and other bioanalytical approaches, this is usually reported using sensitivity, specificity and predictive values. Studies to determine the diagnostic performance of an infectious disease test in terms of sensitivity and specificity are comparatively simple assessments of the proportions of infected and non-infected patients that are correctly and incorrectly identified by the test being evaluated (Baratloo et al., 2015) in comparison to a clinical reference standard (Cohen et al., 2016). The use of sensitivity and specificity in this way reportedly originates from the 1940s as a method to evaluate serological tests for syphilis (Binney et al., 2021). Statistical estimation of confidence (typically 95% confidence intervals) of a particular study are also often included and this may be further simplified to report diagnostic accuracy (Baratloo et al., 2015) as the ability to differentiate those with and without the condition using the test in question. In wider diagnostic practice, a quantitative threshold may be placed on the data to weight the results in favour of being more or less sensitive depending on clinical need. Such receiver-operator characteristics are less common for applied NAAT tests for infectious diseases where such thresholds are less common. This is partly due to the lack of robust quantitative methodology described in the introduction and discussed further below.

A critical problem for routine application of this type of sensitivity/specificity evaluation can be described as the 'gold standard paradox.' The gold standard test is the clinical reference method used to set the bar for any new method to be compared to; for research and development of

NAATs this is invariably another diagnostic test and offers a very simple method for conducting sensitivity and specificity estimation. However, if the new method offers a diagnostic improvement, then this will be penalised as having poor specificity/sensitivity due to the limitations of the gold standard's 'truth'. Furthermore, during a scenario where there is no gold standard clinical reference method, such as when developing a novel method to a new molecular biomarker or an emerging pathogen, this approach causes a problem.

The challenge of the 'gold standard paradox' can be mitigated to some degree by using clinical agreement study comparing percentage agreement between methods and a "state-of-the-art" reference standard (Royal Statistical Society, 2021). In percentage agreement studies the quality of the reference standard is essential and the analytical considerations as outlined above remain important. Furthermore, while sensitivity and specificity assessment using a gold standard has its place, NAATs could arguably also benefit from a more dynamic approach to determine sensitivity, using clinical reference ranges to predict sensitivity as outlined below (and in Fig. 1), than the simple gold standard comparison.

The challenges associated with selecting a gold standard are also an increasing problem for the development of advanced molecular testing that allows for broad range detection of analytes, such as using multiplex syndromic testing or advanced sequencing for metagenomic diagnostics. In many cases it is impractical to demonstrate performance on all potential pathogen sequences (and associated variants) and analytical performance of these approaches will require evaluation of representative sets of reference samples. Ideally these will have been demonstrated to be suitable for assessing the broader pathogen nucleic acids that may be in need for detection; this means the innovation in the routes to support performance evaluation is needed alongside the innovation being applied to develop these advances in molecular IVDs.

5.2. Clinical evaluation, predictive value

Predictive values consider the prevalence (proportion) of the condition in those being tested to estimate the actual impact, in terms of numbers of individuals, effected by false negative and positive results for a given sensitivity and specificity. Studies designed to assess diagnostic performance can, and usually do, use the sensitivity and specificity data to also estimate prevalence to define predictive value (using equation 1 and 2).

1) Positive Predictive Value (PPV)

$$PPV = \frac{true\ positive}{(true\ positive + false\ positive)} \times 100$$

2) Negative Predictive Value (NPV)

$$NPV = \frac{true\ negative}{(true\ negative + false\ negative)} \times 100$$

However, given the fact that these experiments are often deliberately designed and powered to expedite assessment of the clinical performance (requiring specific numbers of positive and negatives to be selected to estimate performance, that do not necessarily reflect the true prevalence) and the prevalence in question may differ with region or over time, this practice is fundamentally flawed (Rutjes et al., 2005). As such approaches may not provide an accurate estimation of prevalence we should stop using equations 1 and 2 to estimate predictive values and instead consider assessing the impact of a given diagnostic test using known, or estimated prevalence (for the population intended to be tested). One way to do this is by using equations 3 and 4 taken from (Tenny and Hoffman).

3) Positive Predictive Value (PPV)

$$PPV = \frac{(sensitivity \times prevalence)}{[(sensitivity \times prevalence) + ((1 - specificity) \times (1 - prevalence))}$$

4) Negative Predictive Value (NPV)

$$NPV = \frac{(specificity\ (1 - prevalence)}{[(specificity \times (1 - prevalence) + ((1 - sensitivity\ ) \times prevalence)]}$$

While equations 3 and 4 may be more useful for estimating predictive value (than equations 1 and 2) they remain limited. This is because, while they allow users to consider prevalence independently of the data used for estimating accuracy, these equations do not consider changes in prevalence that will occur during different stages of an infectious disease outbreak or within different populations (as a result of geography or choice of testing criteria). This also assumes the sensitivity and specificity are fixed properties which we also know may not be the case (Evans et al., 2021). As the prevalence of an infectious disease is usually relatively low, further challenging performance assessment (Holtman et al., 2019), it is the positive predictive value that is often important.

The PPV considers the impact of specificity, in the form of false positives, to rule-in individuals in the population as having the condition being tested. This can be best understood if we imagine using a test on a population that does not have the condition. Therefore, another factor which might influence the prevalence is the indication for testing; for example, the pre-test probability of a test used to 'rule in' COVID-19 in a symptomatic patient may differ greatly from that of a test used to 'rule out' COVID-19 as part of a screening asymptomatic individuals to reduce transmission. In this case, all positive results would be false positives and ruled-in; an analogous situation occurs where prevalence is low; in a situation where a test with a specificity of 99% (1% false positive cases) is used to test a condition with prevalence of 10% in those being tested then around 1 in 10 tested individuals will be incorrectly told they have the condition. This will be more of an issue if prevalence is reduced to 1% in those tested, as using a test of 99% specificity will now result in slightly more false positives than correctly identified individuals. Change in prevalence can occur due to disease dynamics (such as seasonal waves of infection) or equally if testing changes to a different population where prevalence may be lower.

As PPV is affected by the fact that there are usually more people without the condition, it does not take much in terms of a prevalence decrease, for even specific methods to result in more false-positive than-true positive results. Consequently, it is important to consider PPV when deciding how to deploy a diagnostic test. In many situations this was especially relevant during the deployment of COVID-19 diagnostic policy as widespread community testing was encouraged, and in cases mandated, regardless of symptoms, as elaborated above.

5.3. Clinical reference ranges

The discussions in Sections 5 and 6 above questioning the approach or rigour applied when evaluating analytical performance and the importance of accurately defined reference materials may be seen by many as overzealous. Early in the COVID-19 pandemic, some thought that standards would not be necessary for PCR tests because PCR tests are sensitive, the virus level was abundant (and thus easy to detect) and quantification was not required. Similar assumptions are often made in the diagnostic community when using NAATs, especially when non quantitative presence-analysis is being discussed.

The notion that presence/absence testing (i.e. detection without quantification) is simple is not without justification, and the assumption that SARS-CoV-2 RNA detection by PCR could be easy to standardise, as outlined in the preceding paragraph, could well have been valid, depending on the distribution of the analyte (see Fig. 1). The point being made here is that whether a given nucleic acid is clinically abundant (and thus its detection simple to standardise) or not, can be defined

analytically. When designing molecular diagnostic methods for presence/absence testing there is a case to be made that a quantitative analysis be conducted to evaluate the clinical reference ranges to determine the performance criteria (reference ranges and LOD) that are required of the test in question (Fig. 1).

If different tests are used with different LODs and the distribution of the quantities of the analyte being measured spans that region (Fig. 1 A.i and B.i), the different test LODs will score specimens with lower amounts of analyte differently and sensitivity will vary (Evans et al., 2021). If, however, the analyte quantity is always above the LOD of the different tests used (Fig. 1 A.iii and B.iii) then sensitivity will be similar when using different tests. While this may seem glaringly obvious, diagnostic tests are not currently selected to consider this fact.

What this means in terms of clinical sensitivity depends on the condition in question. However, these reference ranges can be used to predict the LOD required for a given clinical sensitivity prior to diagnostic test development and selection. Where analyte quantities are always abundant, then LOD requirement may not need to be as strict (Fig. 1iii), but if clinically low amounts of analyte need to be detected then a more stringent LOD may be needed. TPPs currently report analytical targets (like LOD) independently of clinical targets, yet they can be intrinsically linked with reference ranges and potentially be used to guide the selection of targets. This is especially the case for methods with higher LOD, such as for lateral flow devices, many of which target pathogen protein and have higher LODs.

Reference ranges could also be used to more directly link the analytical sensitivity to the clinical performance. Currently the LOD and the clinical sensitivity (e.g. as outlined in Table 2) are independent of each other and what is proposed above could provide more foresight with an alternative route to just selecting a blanket LOD of 1000 copies/mL as a cautious choice due to lack of information. What we are proposing is that the pathogen nucleic acid reference ranges present within the clinical samples are considered when defining the acceptable LOD, noting that this target may be different during different stages of an outbreak allowing more appropriate deployment of less sensitive tests (Evans et al., 2021).

Reference methods, like dPCR or accurately calibrated qPCR offer a route to quantify the analyte in question and, in combination with methods like massively parallel sequencing, provide wider information on the characteristics of the nucleic acids present within the specimen and could aid in understanding how control materials can be used more accurately in support of routine testing. This could also be used to estimate the reference ranges, explore if and how different genetic targets may differ in abundance, depending on genomic position or size within clinical specimens being tested or control materials used to validate those tests. Importantly, the reference methods used to characterise the specimens may not be practical as a diagnostic solution for reasons including expense, throughput, complexity etc, but they may nevertheless be ideal for accurately characterising the quantities and identity of analytes present within a clinical specimen.

5.4. *Nucleic acid diagnostic threshold*

Quantitative clinical thresholds (or cut-offs) are applied in many examples of clinical chemistry, with measurements of the analyte used to stratify patients into two or more groupings. Examples include prostate-specific antigen (PSA) for prostate cancer, ferritin for iron deficient anaemia (IDA), hemoglobin A1c (HbA1C) for diabetes, cardiac troponin for acute myocardial infarction, and D-Dimer for venous thrombo-embolism. Arguably the most accurate and robust examples of applying thresholds using molecular measurements are not in clinical analysis but in food testing. Food labelling is vital for consumer confidence and food adulteration is a problem which can be detected using molecular methods (Burns et al., 2016). Often these methods are applied to determine defined legislative or agreed quantitative thresholds to label a product as (for example) not containing genetically modified organisms (GMOs). As such, the labelling of a product can draw considerable legal scrutiny and so the methods are conducted with very high degree of accuracy, with clear definitions on route of traceability.

While the above demonstrates quantitative thresholding is possible, with the exception of some bloodborne viruses and examples of minimal residual disease monitoring, it is seldom conducted robustly in applied clinical molecular analysis. This is partly due to the dynamic nature of a disease, challenges over traceability and complexity of multistep procedure (including the taking of the specimen), but also to the challenges associated with making such a measurement reproducible and a wide misunderstanding of the importance of reproducibility for routine clinical use. Establishing a nucleic acid diagnostic threshold may be desirable and, at first, appear to be fairly straightforward once a numerical threshold is agreed; in reality however, this quantitative metric requires considerable support in terms of reference standards to work reproducibly. This was exemplified during COVID-19 as outlined in the following section.

6. Experience from COVID-19

The distribution of SARS-CoV-2 RNA within a population of positive individuals not only varied by over eight orders of magnitude from sample to sample, but the underlying distribution also varied with incidence (Hay et al., 2021). Consequently, when the reproduction number and incidence was high, the amount of SARS-CoV-2 RNA within specimens was much higher. When incidence was lower a larger number of clinically relevant specimens with much lower quantities of viral RNA was observed (Hay et al., 2021). This was because at higher incidence more people in a given population will have been recently infected and have higher RNA quantities (as associated with earlier infection) within the resultant specimens.

Following the discussion in Section 5.3, this meant that methods with higher LODs, of which there were many examples of COVID-19 diagnostic NAATs reported to the FDA (MacKay et al., 2020), would have varied in clinical sensitivity as a result of change in distribution of the RNA quantities ranges with incidence (Evans et al., 2021). This 'sensitivity shift' has implications when diagnostic test results are used at the population level to monitor disease progression or evaluate novel test performance. Under these circumstances, understanding the likely pathogen quantitative distributions (and distributions of the associated RNA and protein analytes) within the population of specimen being examined could aid the deployment of methods during different stages of the COVID-19 pandemic. As similar dynamic ranges of quantity have been reported for influenza A (Van Wesenbeeck et al., 2015) and mpox (Lim et al., 2023), which also vary in incidence, it may be prudent to consider how quantities vary during the different stages of an outbreak.

It may be tempting to conclude that those positive results, missed with a test of higher LOD, that have lower viral burden and are from patients with less severe disease or of lower risk of spreading infection; i.e. the test that is less able to pick up small amounts of virus is actually better at picking up the clinically relevant positive patients with the more sensitive tests actually being less clinically specific. While there may be many situations where this assumption is valid, the clinical relevance of a given nucleic acid amount is highly dependent on the specimen type, as well as sampling procedure. Consequently, pathogen abundance equaling clinical relevance must not be assumed, especially when considering pathogens like SARS-CoV-2 that vary in anatomical location during infection.

While it may be prudent to assume a patient with a highly positive nasal swab may pose a transmission risk, the reverse assumption simply does always not stand as a given. Just because the result from a nasal swab is low (and potentially the LOD of a test format) does not mean there is no transmissible virus elsewhere anatomically; there is evidence that amount of viral nucleic acid does not necessarily correspond with clinical relevance in COVID-19 (Williams et al., 2021) or mpox (Pan et al., 2023a). This further highlights why assuming methods with

higher LODs, which need greater abundance of analyte to work, are more clinically relevant, may be erroneous. Furthermore, a low viral titre from a nasal specimen leading to low or negative result does not provide information on the titre within the lungs, which may actually be high and a major source of transmissibility (Pan et al., 2023b).

The assumptions in the preceding paragraph are also aligned to those outlined in Section 5.4, on the setting of thresholds and the idea that RNA quantities from oral specimens were clinically relevant. During the height of the COVID-19 pandemic the use of the qPCR output unit, the quantification cycle (C_q) or cycle threshold (C_t), was proposed for setting quantitative thresholds to stratify risk (Jefferson et al., 2020; Tom and Mina, 2020), evaluate method performance or even set LOD targets (WHO, 2020). The C_q is determined by setting a fluorescence threshold at a position on the qPCR amplification plot to determine the corresponding cycle (at that position). C_q can vary (by >3 cycles) even when it is set for the same qPCR profile and because the cycles being reported are logarithmic this can amount to orders of magnitude differences in corresponding quantities. C_q represents a crude estimate of nucleic acid quantity that can be used to explore and inform epidemiological observations where large data sets are available; as elegantly demonstrated by Hay et al. (2021). However, as tempting as it is to use as a unit of measure, without calibration it is simply too variable for estimating nucleic acid quantitative thresholds on an individual specimen.

The use of C_q to set thresholds during COVID-19 also reflects a great example of collective amnesia associated within a scientific community. Eminent scientists across the world advocated C_q thresholding for COVID-19 apparently unaware that over 20 years ago a global system was set to support NAAT quantitative analysis (Baylis et al., 2019) due to what was recognised as orders of magnitude quantitative error between laboratories (Fryer et al., 2008). A tangible analogy for the attempted use of C_q to threshold during COVID-19 would be attempting to age children based on their height: for example, if school entry could be based on a child passing the threshold of being over a meter. It is well recognised that children's height is too variable for this to be a useful measure, even though the actual length measurement is accurate. Unlike length measurement, the C_q estimation of viral RNA is in itself not accurate, with >100 fold variation between laboratories being the norm before any actual biological variation is considered (Evans et al., 2021).

The use of thresholding of nucleic acid quantities is applied to stratify and manage patients, such as for chronic myelogenous leukaemia, HIV and HBV but, for this to happen, calibration is required. This removes C_q variation associated with instrument or assay by converting it to a more accurate copy-based unit traceable to a system such as the WHO international standard (using international units [IU] or copies). Using reference measurement systems (such as WHO standards) also allows the copy-based unit to consider differences in the volume of specimen (often standardised to per mL) to be considered which is not possible when using C_q alone. This process also enables the method's linearity to be characterised, adding further accuracy to the measurement. Nucleic acid thresholding, that accounts for that these sources of error, can be used to guide the management of patients with conditions like COVID-19 (Vierbaum et al., 2022), but, for this to be this the method must be validated for this purpose.

An additional factor that COVID-19 placed into the spotlight was how information linked to molecular diagnostic development, research, translation and wider application was reported. NAATs are unique amongst IVD formats as the genetic sequence they target can be 'read' and test developed very quickly using design expertise and synthetic production of oligonucleotides. This should in theory make independent evaluation of procedures more accessible. However, to do this peer reviewed research can be aided in this respect by following specific guidelines (Bustin et al., 2009; Bossuyt et al., 2015), to assist wider application, and manufacturers can provide clear, and appropriate, information on how methods have been validated, verified and accredited.

Possibly the most contentious example of this when considering NAATs is the release of oligonucleotide sequences used to prime the different NAATs. The disclosure of primer sequences has represented a topic of heated debate for some time with strong views on both sides and, in cases, compromises being met (Bustin et al., 2011). Primer sequences are also provided by manufacturers to some, but not all, regulatory authorities. There are certain situations where the need for primer disclosure can be far more strongly argued. The development and application of NAAT-based diagnostic solutions for prevention or response to outbreaks represent such an example. Where outbreak, or potential outbreak, pathogens alter their genetics, such as with SARS-CoV-2 variants of concern or influenza genomic shift and drift, knowledge of the primer sequences could quickly provide regulators and key decision makers with confidence about which IVD solutions are or are not potentially affected by the associated sequence change. If, as we state above, IVDs represent the first response to stopping potential outbreaks, disclosure of primers sequences by IVD manufacturers to immediately remove of the associated ambiguity associated with a new pathogen variant would seem an obvious progressive step forward.

7. Future

There is little doubt that we will see an increased clinical use of molecular bioanalytical techniques including NAATs and sequencing (along with advances in antigen and serological methodologies) in the future. Together these bioanalytical approaches will both increase the analytical information via multiplexed and -omic methodologies, and offer simpler near-patient formats. We are quite simply at the beginning of a potential healthcare revolution that is driven by testing. This is already driving improved cancer treatment, safer prenatal screening and fast pathogen diagnosis. While arguments abound about the diagnostic decisions during COVID-19, one undeniable result is that the public accept testing as a direct route to impact decisions around their own healthcare and behaviour to benefit society. It is now incumbent on the scientific community comprising diagnostic researchers, manufacturers and those tasked with ensuring test quality, to deliver tests that perform with an accuracy that meets this globally developed public trust.

To meet this responsibility the molecular diagnostic sector arguably needs to rethink the approach to molecular diagnostic evaluation commensurate with maximising the potential impact of these highly diverse and dynamic methodologies. Improvements to analytical considerations and how they influence clinical findings will serve to better understand how tests can be used. The process of detection of a pathogen's nucleic acid sequence within a population of specimens may or may not be straightforward, and its increasing abundance may or may not be relevant to clinical severity or transmission. When conducting the research to determine the clinical relevance of the presence, or abundance, of a given pathogen's nucleic acids within a clinical specimen it would be prudent to accurately define the analytical performance of the methods in question. This in turn requires improved support in terms of the availability and application of material and methodological standards, and how they should be best used to aid in the setting and meeting of performance targets.

COVID-19 demonstrated the important role for bioanalytical techniques in identifying and diagnosing the offending pathogen. Without advanced sequencing and PCR, our response to COVID-19 would have been very different. If diagnostics are elevated in importance to be more commensurate with therapeutics and vaccines, then we will have a greater chance to impact in the next pandemic threat as well as in delivering the wider medical life science advancements many predict.

CRediT authorship contribution statement

Jim F. Huggett: Conceptualization, Funding acquisition, Investigation, Methodology, Writing – original draft. **Denise M. O'Sullivan:** Conceptualization, Formal analysis, Investigation, Supervision, Writing – review & editing. **Simon Cowen:** Conceptualization, Data curation, Formal analysis, Software, Writing – review & editing. **Megan H.**

Cleveland: Conceptualization, Investigation, Writing – review & editing. **Kerrie Davies:** Conceptualization, Investigation, Visualization, Writing – review & editing. **Kathryn Harris:** Conceptualization, Investigation, Validation, Writing – review & editing. **Jacob Moran-Gilad:** Conceptualization, Investigation, Validation, Writing – original draft. **Amanda Winter:** Conceptualization, Methodology, Writing – original draft. **Julian Braybrook:** Supervision, Validation, Writing – review & editing. **Michael Messenger:** Conceptualization, Methodology, Writing – original draft.

Acknowledgements

This work was supported by the UK National Measurement System and the European Metrology Program for Innovation and Research (EMPIR) joint research project [18HLT03] "SEPTIMET," which has received funding from the EMPIR program cofinanced by the Participating States and the European Union's Horizon 2020 research and innovation program.

Kerrie Davies is supported in part by the National Institute for Health and Care Research (NIHR) Leeds Biomedical Research Centre (BRC) (NIHR213331). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care."

Caroline Wilson, The Newcastle upon Tyne Hospitals NHS Foundation Trust, North East Innovation Laboratory, The Biosphere, Drayman's Way, Newcastle upon Tyne, NE4 5BX.

Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Commerce. Certain commercial software, instruments, and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose.

All figures were created with BioRender.com.

References

Armbruster, D.A., Pry, T., 2008. Limit of blank, limit of detection and limit of quantitation [eng]. Clin. Biochem. Rev. 29 (Suppl. 1), S49–S52.

International Organization of Standardization 2020. In Vitro Diagnostic Test Systems Nucleic Acid Amplification-Based Examination Procedures for Detection and Identification of Microbial Pathogens Laboratory Quality Practice Guide, ISO 17822.

Baratloo, A., Hosseini, M., Negida, A., 2015. El Ashal G. Part 1: simple definition and calculation of accuracy, sensitivity and specificity. Emerg (Tehran) 3 (2), 48–49. Spring.

Baylis, S.A., Wallace, P., McCulloch, E., Niesters, H.G.M., Nübling, C.M., 2019. Standardization of nucleic acid tests: the approach of the World Health Organization [eng]. J. Clin. Microbiol. 57, 1. https://doi.org/10.1128/jcm.01056-18.

Binney, N., Hyde, C., Bossuyt, P.M., 2021. On the origin of sensitivity and specificity. Ann. Intern. Med. 174 (3), 401–407. https://doi.org/10.7326/m20-5028.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L., Lijmer, J.G., et al., 2015. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Clin. Chem. 61 (12), 1446–1452. https://doi.org/10.1373/clinchem.2015.246280. Epub 20151028 as.

Burns, M., Wiseman, G., Knight, A., Bramley, P., Foster, L., Rollinson, S., Damant, A., Primrose, S., 2016. Measurement issues associated with quantitative molecular biology analysis of complex food matrices for the detection of food fraud [eng]. Analyst 141 (1), 45–61. https://doi.org/10.1039/c5an01392e. Epub 20151203 as.

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., et al., 2009. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments [eng]. Clin. Chem. 55 (4), 611–622. Epub 2009/02/28 as doi: clinchem.2008.112797 [pii]10.1373/clinchem.2008.112797 [doi].

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., et al., 2011. Primer sequence disclosure: a clarification of the MIQE guidelines. Clin. Chem. 57 (6), 919–921. https://doi.org/10.1373/clinchem.2011.162958. Epub 20110318 as.

Cohen, J.F., Korevaar, D.A., Altman, D.G., Bruns, D.E., Gatsonis, C.A., Hooft, L., Irwig, L., et al., 2016. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration [eng]. BMJ Open 6, 11. https://doi.org/10.1136/bmjopen-2016-012799. Epub 20161114 as.

European Union. Commission Implementing Regulation (EU) 2022/1107 of 4 July 2022 laying down common specifications for certain class D in vitro diagnostic medical devices in accordance with Regulation (EU) 2017/746 of the European Parliament and of the Council (Text with EEA relevance) https://eur-lex.europa.eu/eli/reg_impl/2022/1107/oj. (Accessed..

dMIQE, Group., Huggett, J.F., 2020. The digital MIQE guidelines update: minimum information for publication of quantitative digital PCR experiments for 2020. Clin. Chem. 66 (8), 1012–1029. https://doi.org/10.1093/clinchem/hvaa125.

Evans, D., Cowen, S., Kammel, M., O'Sullivan, D.M., Stewart, G., Grunert, H.P., Moran-Gilad, J., et al., 2021. The dangers of using Cq to quantify nucleic acid in biological samples: a lesson from COVID-19 [eng]. Clin. Chem. 68 (1), 153–162. https://doi.org/10.1093/clinchem/hvab219.

FDA, 2021. Supplemental template for Developers of molecular and antigen diagnostic COVID-19 tests for screening with serial testing. https://www.fda.gov/media/146695/download.

Fryer, J.F., Baylis, S.A., Gottlieb, A.L., Ferguson, M., Vincini, G.A., Bevan, V.M., Carman, W.F., Minor, P.D., 2008. Development of working reference materials for clinical virology [eng]. J. Clin. Virol. 43 (4), 367–371. Epub 2008/10/01 as doi: S1386-6532(08)00298-9 [pii]10.1016/j.jcv.2008.08.011.

G7, 2021. 1OO DAYS MISSION to respond to future pandemic threats. https://www.gov.uk/government/publications/100-days-mission-to-respond-to-future-pandemic-threats.G7. Summit 2021.

Hay, J.A., Kennedy-Shaffer, L., Kanjilal, S., Lennon, N.J., Gabriel, S.B., Lipsitch, M., Mina, M.J., 2021. Estimating epidemiologic dynamics from cross-sectional viral load distributions [eng]. Science 373, 6552. https://doi.org/10.1126/science.abh0635. Epub 20210603 as.

Hijano, D.R., Brazelton de Cardenas, J., Maron, G., Garner, C.D., Ferrolino, J.A., Dallas, R.H., Gu, Z., Hayden, R.T., 2019. Clinical correlation of influenza and respiratory syncytial virus load measured by digital PCR [eng]. PLoS One 14 (9), e0220908. https://doi.org/10.1371/journal.pone.0220908. Epub 20190903 as.

Holtman, G.A., Berger, M.Y., Burger, H., Deeks, J.J., Donner-Banzhoff, N., Fanshawe, T. R., Koshiaris, C., et al., 2019. Development of practical recommendations for diagnostic accuracy studies in low-prevalence situations [eng]. J. Clin. Epidemiol. 114, 38–48. https://doi.org/10.1016/j.jclinepi.2019.05.018. Epub 20190528 as.

Huggett, J.F., Benes, V., Bustin, S.A., Garson, J.A., Harris, K., Kammel, M., Kubista, M., et al., 2020. Cautionary note on contamination of reagents used for molecular detection of SARS-CoV-2 [eng]. Clin. Chem. https://doi.org/10.1093/clinchem/hvaa214. Epub 2020/09/08 as.

Huggett, J.F., Moran-Gilad, J., Lee, J.E., 2021. COVID-19 new diagnostics development: novel detection methods for SARS-CoV-2 infection and considerations for their translation to routine use. Curr. Opin. Pulm. Med. 27 (3), 155–162. https://doi.org/10.1097/MCP.0000000000000768.

Jefferson, T., Spencer, E.A., Brassey, J., Heneghan, C., 2020. Viral cultures for coronavirus disease 2019 infectivity assessment: a systematic review. Clin. Infect. Dis. 73 (11), e3884. https://doi.org/10.1093/cid/ciaa1764.

Kim, D., Lee, J.Y., Yang, J.S., Kim, J.W., Kim, V.N., Chang, H., 2020. The architecture of SARS-CoV-2 transcriptome [eng]. Cell 181 (4), 914. https://doi.org/10.1016/j.cell.2020.04.011, 21.e10. Epub 20200423 as.

Lim, C.K., McKenzie, C., Deerain, J., Chow, E.P.F., Towns, J., Chen, M.Y., Fairley, C.K., et al., 2023. Correlation between monkeypox viral load and infectious virus in clinical specimens [eng]. J. Clin. Virol. 161, 105421 https://doi.org/10.1016/j.jcv.2023.105421. Epub 20230303 as.

MacKay, M.J., Hooker, A.C., Afshinnekoo, E., Salit, M., Kelly, J., Feldstein, J.V., Haft, N., et al., 2020. The COVID-19 XPRIZE and the need for scalable, fast, and widespread testing [eng]. Nat. Biotechnol. 38 (9), 1021–1024. https://doi.org/10.1038/s41587-020-0655-4. Epub 2020/08/21 as.

Mercer, T., Almond, N., Crone, M.A., Chain, P.S.G., Deshpande, A., Eveleigh, D., Freemont, P., et al., 2022. The Coronavirus Standards Working Group's roadmap for improved population testing [eng]. Nat. Biotechnol. 40 (11), 1563–1568. https://doi.org/10.1038/s41587-022-01538-1.

Mercier, B., Gaucher, C., Feugeas, O., Mazurier, C., 1990. Direct PCR from whole blood, without DNA extraction [eng]. Nucleic Acids Res. 18 (19), 5908. https://doi.org/10.1093/nar/18.19.5908.

MHRA, 2023. Target product profile: laboratory-based SARS-CoV-2 viral detection tests. https://www.gov.uk/government/publications/how-tests-and-testing-kits-for-coronavirus-covid-19-work/target-product-profile-laboratory-based-sars-cov-2-viral-detection-tests.

Nixon, G., Garson, J.A., Grant, P., Nastouli, E., Foy, C.A., Huggett, J.F., 2014. Comparative study of sensitivity, linearity, and resistance to inhibition of digital and nondigital polymerase chain reaction and loop mediated isothermal amplification assays for quantification of human cytomegalovirus [eng]. Anal. Chem. 86 (9), 4387–4394. https://doi.org/10.1021/ac500208w. Epub 20140422 as.

Page, M., Almond, N., Rose, N.J., Schneider, C.K., 2020. Diagnostics and the coronavirus: don't let the standards slip [eng]. Nat. Biotechnol. 38 (6), 673–674. https://doi.org/10.1038/s41587-020-0558-4. Epub 2020/05/24 as.

Pan, D., Atkinson, B., Decker, J., Williams, C.M., Nazareth, J., Martin, C.A., Bird, P., et al., 2023a. Concomitant, consecutive, self-obtained facemask and swab samples from exhaled breath, pox lesions, nasopharynx and the face in patients recovering from mpox - a longitudinal sampling study [eng]. J. Infect. 9 https://doi.org/10.1016/j.jinf.2023.05.007. Epub 20230509 as.

Pan, D., Williams, C.M., Decker, J., Fletcher, E., Sze, S., Assadi, S., Haigh, R., et al., 2023b. Exhaled SARS-CoV-2 RNA viral load kinetics measured by facemask sampling associates with household transmission [eng]. Clin. Microbiol. Infect. 29 (2) https://doi.org/10.1016/j.cmi.2022.07.005, 254.e1–.e6. Epub 20220714 as.

Peeling, R.W., Heymann, D.L., Teo, Y.Y., Garcia, P.J., 2022. Diagnostics for COVID-19: moving from pandemic response to control [eng]. Lancet 399 (10326), 757–768. https://doi.org/10.1016/s0140-6736(21)02346-1. Epub 20211220 as.

Royal Statistical Society, 2021. Royal statistical society diagnostic tests working group report. https://rss.org.uk/RSS/media/File-library/Policy/2021/RSS-Diagnostic-tests-report-FINAL.pdf.

Rutjes, A.W., Reitsma, J.B., Vandenbroucke, J.P., Glas, A.S., Bossuyt, P.M., 2005. Case-control and two-gate designs in diagnostic accuracy studies [eng]. Clin. Chem. 51 (8), 1335–1341. https://doi.org/10.1373/clinchem.2005.048595. Epub 20050616 as.

Sandberg, S., Fraser, C.G., Horvath, A.R., Jansen, R., Jones, G., Oosterhuis, W., Petersen, P.H., et al., 2015. Defining analytical performance specifications: consensus statement from the 1st strategic conference of the European federation of clinical chemistry and laboratory medicine. Clin. Chem. Lab. Med. 53 (6), 833–835. https://doi.org/10.1515/cclm-2015-0067 [eng].

Tenny S, Hoffman M. Prevalence. In: StatPearls 2023;Treasure Island (FL): StatPearls Publishing https://www.ncbi.nlm.nih.gov/books/NBK430867/..

Tom, M.R., Mina, M.J., 2020. To interpret the SARS-CoV-2 test, consider the cycle threshold value. Clin. Infect. Dis. https://doi.org/10.1093/cid/ciaa619. Epub 2020/05/22 as.

Van Wesenbeeck, L., D'Haese, D., Tolboom, J., Meeuws, H., Dwyer, D.E., Holmes, M., Ison, M.G., et al., 2015. A downward trend of the ratio of influenza RNA copy number to infectious viral titer in hospitalized influenza A-infected patients [eng]. Open Forum Infect. Dis. 2 (4), ofv166 https://doi.org/10.1093/ofid/ofv166. Epub 20151103 as.

Vierbaum, L., Wojtalewicz, N., Grunert, H.P., Lindig, V., Duehring, U., Drosten, C., Corman, V., et al., 2022. RNA reference materials with defined viral RNA loads of SARS-CoV-2-A useful tool towards a better PCR assay harmonization [eng]. PLoS One 17, 1. https://doi.org/10.1371/journal.pone.0262656. Epub 20220120 as.

WHO, 2020. Target Product Profiles for Priority Diagnostics to Support Response to the COVID-19 Pandemic, 1.0.

WHO, 2023. Target Product Profiles for Tests Used for Mpox (Monkeypox) Diagnosis.

WHO. Technical Specifications Series https://extranet.who.int/prequal/vitro-diagnostics/technical-specifications-series.

WHO. Target Product Profile Directory. https://www.who.int/our-work/science-division/research-for-health/target-product-profile-directory.

Williams, C.M., Pan, D., Decker, J., Wisniewska, A., Fletcher, E., Sze, S., Assadi, S., et al., 2021. Exhaled SARS-CoV-2 quantified by face-mask sampling in hospitalised patients with COVID-19 [eng]. J. Infect. 82 (6), 253–259. https://doi.org/10.1016/j.jinf.2021.03.018. Epub 20210324 as.

Wu, X., Wang, Z., He, Z., Li, Y., Wu, Y., Wang, H., Liu, Y., et al., 2021. A follow-up study shows that recovered patients with re-positive PCR test in Wuhan may not be infectious [eng]. BMC Med. 19 (1), 77. https://doi.org/10.1186/s12916-021-01954-1. Epub 20210315 as.

Yoo, H.B., Park, S.R., Dong, L., Wang, J., Sui, Z., Pavšič, J., Milavec, M., et al., 2016. International comparison of enumeration-based quantification of DNA copy-concentration using flow cytometric counting and digital polymerase chain reaction [eng]. Anal. Chem. 88 (24), 12169–12176. https://doi.org/10.1021/acs.analchem.6b03076. Epub 20161130 as.