Guidance for surveillance of SARS-CoV-2 variants

Interim guidance

9 August 2021

World Health Organization

Key points

- The public health risks of known and emerging variants of interest (VOIs) and variants of concern (VOCs) can be categorized into five main domains: increased transmissibility; a more severe clinical course; failure to be detected by diagnostic assays; escape to natural or vaccine-derived immunity and decreased susceptibility to therapeutics.
- Routine genetic sequencing is critical to follow the emergence and impact of VOIs and VOCs. Countries with limited capacity to perform sequencing are strongly encouraged to take steps to facilitate access to regional and international sequencing partnerships or increase their capacity through existing sequencing systems or laboratory networks.
- Sampling for genetic sequencing should consider all the following subsets, as feasible:
  - randomized samples, representative of geographic and demographic distribution of SARS-Cov-2 infections
  - targeted sampling focusing on specific subsets of cases associated with public health risks: diagnostic failures, vaccinated cases, reinfections, immunocompromised cases
  - outbreaks, alerts or other unusual events.
- Unexpected trends or signals from routine epidemiological surveillance (or other sources) such as increasing trends in epidemic course, with high impact on public health, can be an indication of a potential VOI or VOC.
- All reported sequences should be associated with a minimum set of linked information, called metadata, and include core details. If possible, descriptive metadata and metadata for characterization should be included.
- A combination of laboratory science, clinical manifestations and detailed epidemiologic investigations is required to accurately and rapidly characterize public health risks of SARS-CoV-2 variants.
- Prompt sharing of information around SARS-CoV-2 variant genomic sequences in public databases is integral to the global understanding and control of SARS-CoV-2.

Purpose of the document

This document aims to describe a minimum set of surveillance activities recommended at the national level to detect and monitor the relative prevalence of SARS-CoV-2 variants and outline a set of activities for the characterization and assessment of risk posed by these variants. A set of indicators is also provided to standardize monitoring and public reporting of variant circulation.

The document is primarily intended for national and sub-national public health authorities and partners who support implementation of surveillance for SARS-CoV-2 variants. Additional guidance has been published for laboratory stakeholders on diagnostic testing for SARS-CoV-2 and sequencing for public health goals, alongside an implementation guide for SARS-CoV-2 sequencing.

Background

SARS-CoV-2 is an enveloped, positive-sense single-stranded RNA virus with a 30-kilobase genome, which, like all viruses, accumulates nucleotide mutations over time. These mutations result in the formation of distinct viral lineages. Since its characterization (1), genomic sequencing of SARS-CoV-2 has been conducted to identify mutations and any corresponding amino acid substitutions. Although the emergence of these new variants is expected and most have no impact on viral behaviour, some mutations may produce changes in phenotype.

The public health risks of known and emerging variants can be broadly categorized into five main domains:

- increased transmissibility, due to increased viral shedding, binding affinity for host cells, or stability of the virus
- atypical clinical course (e.g. increased severity, atypical signs and symptoms)
- diagnostic failure: decreased performance of some laboratory diagnostics, particularly molecular assays such as nucleic acid amplification testing (NAAT) (2) and approved antigen-detection rapid diagnostic tests
- decreased effectiveness of natural and vaccine-derived immunity: the ability of the variant to partially evade the host antibody response and, potentially, increase the likelihood of reinfection or vaccine breakthrough
- decreased susceptibility to therapeutics: the potential ability of a new variant to evade an antibody therapy is a cause for concern (3) and has led to changes in recommendations on the use of some therapeutics.

The World Health Organization (WHO) classifies "variants of interest" (VOIs) or "variants of concern" (VOCs) according to the global impact of these factors. As of 9 July 2021,WHO had designated seven VOIs and four VOCs (4).

Global genomic surveillance for SARS-CoV-2 is a critical public health function, with the primary objective to inform national and global decisions around public health and social measures (PHSMs), diagnostics, therapeutics, and vaccination. Surveillance of variants can be done through genomic surveillance as well as through detection of epidemiological signals and unexpected trends. These two strands of evidence should be brought together in a timely fashion to provide a broad understanding of viral evolution and its potential impact on disease control, in order to guide public health response.

Despite the concerning phenotypes of known VOIs and VOCs, WHO continues to recommend the implementation and adjustment of PHSMs to control transmission, as described in existing WHO guidance. However, close monitoring of the impact of current variants on the efficacy of PHSMs is needed.

Guidance on diagnostic testing for SARS-CoV-2 can be found here, and specific guidance on the use of antigen-detection rapid diagnostics tests can be found here.

Evidence that vaccines may be less protective against a specific variant may be suggested by genomic and structural biology studies, animal studies and in-vitro neutralization testing. Lower effectiveness of a vaccine in protecting against infection and disease from a variant in humans, however, provides the strongest evidence. Epidemiological data on vaccine performance against new variants will primarily come from observational studies of vaccine effectiveness (VE); please see the Addendum to Evaluation of COVID-19 vaccine effectiveness.

More variants will likely continue to arise as transmission continues, and they may be subject to selective pressures from natural immunity, vaccine use and therapeutics.

Methodology

This WHO interim guidance was written by WHO and the United States Centers for Disease Control and Prevention (US CDC) in consultation with the Africa Centers for Disease Control and Prevention (ACDC), the European Center for Disease Control and Prevention (ECDC), with additional feedback from expert advisory groups, such as WHO's Epidemiology Technical Advisory Group. The guidance is based on a review of emerging evidence on variant epidemiology and characterization methods, covering all regions, using English language search engines. Search topics included: public health driven sampling strategy, genomic sequencing, phylogeny, genomic epidemiology and surveillance methods, genomic metadata and database management, genomic database analysis for public health purposes, as well as specific characterization evidence for specific mutations and variants of concern and interest. Evidence was synthesized following the thematic sections of the guidance. Additional references were provided by experts, and existing guidance documents from WHO and other partners were also referenced; see table below. This document is subject to updates as new evidence and methodologies on variant investigation emerge.

1. Surveillance for SARS-CoV-2 variants

1.1. Building capacity for genomic sequencing

Capacities for SARS-CoV-2 sequencing activities have expanded considerably as the pandemic has evolved. However, sequencing capacity varies significantly within and between countries. As a result, the amount of genetic sequence data (GSD), quality of metadata accompanying the GSD and the length of time from specimen collection to sequencing and reporting differ widely across countries. To help address this challenge, on 8 January 2021, WHO published two interim guidance documents on SARS-CoV-2 genomic sequencing to track the geographic spread of the virus over time and ensure that mutations with the potential to influence transmissibility, pathogenicity and medical countermeasures are identified and assessed quickly: SARS-CoV-2 genomic sequencing for public health goals and Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health. The United States Centers for Disease Control and Prevention (US CDC) has also released a COVID-19 Genomic Epidemiology Toolkit (6).

Countries with limited capacity to perform sequencing are strongly encouraged to take steps to facilitate access to existing regional and international sequencing networks and partnerships. Countries may also choose to increase sequencing capacity through existing surveillance systems with sequencing capacity such as the **Global Influenza Surveillance and Response System** (GISRS) or existing regional networks. The Shipping Fund Project is designed to support the transport of samples for sequencing and data sharing. Collaborations with experienced laboratories and other potential partners may include public health departments, non-profits, academic centres or commercial entities. In addition to whole genomic sequencing, countries can also screen for known mutations utilizing targeted RT-PCR-based mutation detection assays; countries are encouraged to establish clear procedures for their use.

1.2. Variant definitions

WHO has released working case definitions for SARS-CoV-2 variants of interest and variants of concern. These may be updated regularly. Please refer to the WHO variant webpage for latest definitions, and list of latest VOIs and VOCs.

1.3. Variant alert triggers

Unexpected trends or signals from routine epidemiological surveillance (or other sources) indicating an increased impact of the course of the pandemic on public health can be an indication of a potential VOI or VOC.

1.3.1. Routine Epidemiological surveillance

WHO recommends the following minimum set of variables to be included in weekly epidemiological surveillance.

- Number of confirmed cases
- Number of probable cases
- Number of confirmed deaths
- Number of probable deaths
- Number of individuals hospitalized (confirmed and probable)
- Number discharged (confirmed and probable)
- Number of health care workers infected (confirmed + probable) as a subset of total cases count
- Number of health care workers who died due to COVID-19 (confirmed + probable) as a subset of total death count
- Number of persons tested
- Number of persons tested by PCR
- Confirmed + probable cases by age group and sex (see below)
- Confirmed + probable deaths by age group and sex (see below)
- Transmission classification

The following age categories (in years) are recommended: 0-4, 5-9, 10-14, 15-19, 20-29, 30-39, 40-49, 50-59, 60-64, 65-69, 70-74, 75-79, and >80.

In addition to these variables, the monitoring of ICU occupancy and vaccination coverage of targeted population subgroups can enhance routine surveillance for alerts.

Weekly monitoring of epidemiological indicators at a high geographical granularity allows for timely detection of any departure from trends, or unexpected signals. This allows for early targeting of investigation and sampling for sequencing. Analysis should take into account public health and social measures, stringency index (7) and any other parameters that could impact transmission (e.g. mass gatherings).

Table 1 Examples of disease surveillance indicators, alert and trigger thresholds

Indicators	Alert
Cases	Increase / departure from trend
Age-disaggregated cases	Increase in specific age groups (under 18, under 65; to be determined locally)
Cases among health and care workers	Increase / departure from trend
Case fatality ratio	Increase / departure from trend
Age disaggregated deaths	Increase in specific age groups
Hospitalizations/ICU admissions or bed occupancy rate	Increase in specific age groups
Test positivity rate	Increase / departure from trend

These triggers and thresholds should be adapted to local situations, investigation capacity and desired sensitivity.

If routine surveillance systems are not in place to monitor hospital or ICU admissions or bed capacity, demand for oxygen and ventilators may indicate a surge in severe illness, which may or may not be driven by an emerging variant with increased virulence. Such indicators can be followed with joint monitoring from pharmaceutical and biomedical supplies providers.

Similarly, increases in transmission beyond which might be expected given population levels of immunity also warrant further investigation. For example, sustained community transmission in areas where vaccination coverage is high or there are high levels of past infection may indicate the presence of a variant able to evade the immune response. Please see guidance on vaccine effectiveness in the context of new SARS-CoV-2 variants.

If robust case investigation and contact tracing protocols are in place, an increasing proportion of contacts that become cases (i.e. unexpectedly high secondary attack rate in comparison to studies in similar settings, such as households) might provide a similar signal.

Trends in mortality at the lowest available administrative level can reveal an increase in death rate in particular populations, and the case fatality ratio (CFR) may be estimated if case-based surveillance data covering the same time period and geographic region are also available (see Case Fatality Ratio scientific brief). Increases in CFR may warrant further investigation through genomic characterization, although trends in mortality are unlikely to reveal a variant with higher severity unless there is a drastic change in CFR. A decoupling between mortality and incidence trends (i.e., mortality higher than expected for a given incidence) could also be an indicator in increased disease severity.

1.3.2. Event-based surveillance

Reports of rapidly spreading outbreaks in healthcare facilities or communities might raise the concern that these events are due to a variant that spreads more easily from person to person. Similar reports from populations expected to have a high level of immunity (due to prior infections or high vaccination coverage) may indicate the presence of a variant able to evade the immune response.

Outbreaks causing unexpectedly high levels of morbidity and mortality (otherwise unexplained by the demographics and underlying conditions of the affected population, clinical case management or hospitalization capacity, medical supply shortage or other factors) may be due to a variant causing more severe disease.

Depending on capacity, such reports may trigger a field investigation. Specimens collected during such investigations may warrant prioritization for sequencing.

Reports of respiratory illness clusters meeting the suspect or probable COVID-19 case definition but testing negative for SARS-CoV-2 and without an alternative clinical diagnosis may also warrant investigation.

1.3.3. Environmental surveillance

If existing surveillance systems are in place to monitor wastewater for SARS-CoV-2 RNA, these might be leveraged for variant surveillance. Viral RNA can be sequenced directly from wastewater and may provide an early glimpse into whether a known VOC is being transmitted. There are several real-world examples of genomic sequencing of SARS-CoV-2 in wastewater revealing variants of concern (8,9),(10), but the temporal and quantitative association with community transmission requires further study (11).

1.4. Sampling strategies

Sampling strategies will vary depending on national variant surveillance objectives. Primary objectives may include:

a) detecting variants circulating at low levels
b) monitoring the relative prevalence of variants across time and geographic areas
c) investigating particular cases of public health interest.

Broadly, objectives (a) and (b) can be met through routine surveillance of a randomized sample. Objective (c) requires a targeted sample.

For countries with high sequencing capacity, priority objectives should be a), detecting variants; and b), monitoring relative prevalence of variants. Countries with low sequencing capacity should focus on b), monitoring relative prevalence of variants.

1.4.1. Representative sampling for routine surveillance

Randomized representative sampling can be defined as a selection of a subset of a given target population, representative of the target population situation. Criteria accounting for the representative distribution of a sample should include at least age, sex, clinical spectrum, and geographical distribution.

Key considerations of surveillance systems outlined in other WHO documents (e.g. GISRS guidance) remain relevant, particularly: systematic sample collection, geographically relevant and sustainable sampling at regular frequency; sampling a representative population; and providing timely sample sequencing and analysis.

When conducting genomic surveillance, it is important to consider the time interval between infection and availability of sequence data. Contributors to a lack of timeliness include lags between sample collection and receipt of specimens by the sequencing laboratory; laboratory processing time; bioinformatics analysis; and time needed to provide data to public health authorities or to post sequence data on public databases. Efforts should be made to increase the timeliness of these actions at every stage. Routine collection of surveillance samples over a fixed, repeating time interval will ensure lagging data are regularly updated. Because the relative prevalence of variants can change rapidly, regular sample collection is recommended, preferably on a weekly basis. This also allows for time series to offer highly dynamic representativity.

The methods of selecting a representative sample may vary by country and draw on local surveillance systems, whether routine or sentinel, such as the GISRS sentinel sites network for influenza-like illness/ severe acute respiratory infection (ILI/SARI). Because incidence might fluctuate rapidly, sampling a fixed number of cases (as opposed to a fixed proportion of cases) may be more logistically feasible for submitting and sequencing laboratories to predict resource requirements and standardize protocols.

Table 2: Sensitivity and specificity of sequencing strategies

	Pros	Cons
1- Randomized representative sampling	High sensitivity	Large sample size: capacity challenge
2- Fixed sample from sentinel sites	Operationally practical; if stable, can allow to follow trend of circulating variants	Low sensitivity Low representativeness (geographical, population based)

1.4.1.1. Sampling methodologies

Sample size calculations assume that specimens are randomly sampled and thereby likely to be representative, assuming that positive samples are themselves a true representation of underlying infection rates. If diagnosed cases are a representative sample of all COVID-19 cases because diagnostic coverage is equally distributed across the country, an unadjusted sample of positive specimens identified through the clinical system may be sufficient to ensure reasonably representativeness.

However, in many countries, diagnostic coverage is uneven because of disparities in access to healthcare and diagnostics or extensive use of contact tracing to identify cases. If diagnostic coverage is not equally distributed, weighting the sample may partially adjust for this. This can be achieved by asking areas with lower diagnostic coverage to submit a higher proportion of specimens than areas with ready diagnostic access.

Options for representative specimen collection may include systematic sampling (sample selection at regular intervals) and random sampling (sample selection generated randomly). The choice of methods should be validated by comparing the distribution of representativeness criteria (e.g. age, sex, clinical spectrum, and geographical distribution) across the sample.

If a truly representative sample is difficult to obtain, sentinel surveillance from sites already enrolled in ILI, ARI or SARI surveillance might provide a valuable platform. Collecting a standard number of specimens from sentinel sites, rather than aiming to be 'representative' geographically, may provide greater stability and improve the quality of specimens and associated metadata, ensuring comparability over time to allow for monitoring of trends. However, depending on existing sentinel sites, this strategy could lead to biased estimations of the relative prevalence of variants and exclusion of some populations or settings.

The sampling methodology should be documented and taken into account during data analysis and interpretation.

1.4.1.2. Sampling size calculations

- **Randomized representative sampling**

Various sample size calculators (12,13) can help to refine the number of specimens from a representative sample that need to undergo genomic sequencing to detect variants circulating at low levels with a specified level of confidence. Given that sequencing capacity is highly variable across countries, and achievable sample sizes may be highly dependent on capacity, it is possible to use these same sample size calculators to 'back-calculate' the level of confidence and precision in available sequence data.

The ECDC has released detailed guidance on sample size calculations to detect and monitor the proportion of variants circulating at low levels, inclusive of tables showing the required sample size under various situations and parameters and the underlying equations to facilitate replication. Considerations when identifying a sample include:

- level of precision/sensitivity to detect
    - A variant circulating at a low level (e.g. 1%) will require a larger sample than that to detect a variant circulating at a higher level.
    - The ability to detect a change in the relative prevalence of a variant from 2.5% to 5% will require a larger sample than that required to detect a change from 2.5% to 10%.
- level of confidence required (e.g. 95% confidence)
- level of transmission within the country (a larger sample will be required when the incidence is high and there are many people with SARS-CoV-2 infection)
- unit of sampling time (regular, routine sampling weekly, every two weeks, or every month) is required, because the relative prevalence of lineages can change rapidly.

The required sensitivity to detect variants circulating at low levels, changes in the relative prevalence of variant lineages and the level of confidence of the surveillance findings, are country-level decisions. In general, for public health purposes, the sensitivity to detect variants circulating at low levels may be the primary driver of sample size decisions, because the public health significance of detecting a variant that was not detected before may be higher than detecting a modest change in the relative prevalence of a given lineage. Additionally, estimates of the sample size needed to monitor relative prevalence are complicated by the number of different lineages in local circulation.

Table 3: *Sample sizes required to detect a significant change (at 95% confidence) of relative prevalence*

Weekly number of SARS-CoV-2 detections	Sample size based on the difference in the proportion of a certain variant, from one week to another	
	From 2.5% to 5%	**From 2.5% to 10%**
>100,000	725	129
10,001–100,000	705–720	129
5,001–10,000	676	128
2,501–5,000	634	126
1,000–2,500	563	123
500–1,000	421	115
<500	296	103

As outlined above, identifying and sequencing a truly random sample is difficult. However, if biases are well understood, adjustment may be possible once sequence results are available, making it possible to provide less biased prevalence estimates. Additionally, given the inevitable delay between specimen collection and availability of sequence results, modelling approaches can project a current state of relative lineage prevalence based on the available sequence data and lineage growth rates, see CDC MMWR (15) and Galloway et al. on the emergence of Alpha (B1.1.7) (16).

- **Fixed sample sizes**

In countries with minimal laboratory capacity, sequencing a minimum of 15 specimens per week from sentinel sites provides a baseline on which to build (WHO GISRS 2021). Africa CDC and the Pathogen Genome Initiative (PGI) network aim to collect a random sample of at least 50 positive specimens from each country per week, with the goal of establishing a sustainable, routine sampling frame for African countries (14), whereas the WHO Regional Office for the Americas/Pan American Health Organization (PAHO) recommends that countries sequence at least 50 positive specimens a month. This is roughly equivalent to detection of at least one sample of a variant that has 5% prevalence for the determined sampling period. If the sample size is fixed, the confidence level of not detecting a specific variant can be 'back-calculated' (12).

1.4.2. Targeted sampling

Targeted sequencing of specimens with a higher pre-test probability of being a VOI or VOC might be beneficial in addition to the above strategies.

Potential triggers for targeted sequencing for surveillance include (see section 3.3):

- specimen-level characteristics [e.g. genomic sequencing based on results from screening assays such as PCR-based single nucleotide polymorphism (SNP) detection assays]
- individual-level characteristics (e.g. clinical characteristics; immunocompromised patients and selective sequencing of vaccine breakthrough)
- environmental characteristics (e.g. evidence of variant sequences from wastewater surveillance).

1.4.2.1. Specimen-level characteristics

A range of RT-PCR primers and probes specific for mutations common to VOCs are now available (17) (18). These assays rely on the detection of single or multiple single nucleotide polymorphisms (SNPs) that are characteristic of specific lineages or shared across multiple lineages and that are often believed to contribute to a phenotypic change. However, these mutations may also be present in non-VOCs, so verification by genomic sequencing is needed for definitive lineage assignment.

PCR-based approaches followed by whole genome sequencing (WGS) have several advantages. First, RT-PCR is more readily available and less resource intensive than sequencing and can therefore be done across a wider geographical area at higher volume. Second, RT-PCR results may provide information more rapidly than WGS, which often requires specimen transport to a reference laboratory. Third, if applied to a larger sample size than WGS, PCR pre-screening may allow detection of a lineage circulating at a low relative frequency.

However, restricting sequencing to samples that have been pre-screened using the SNP RT-PCR tests has limitations. First, PCR assays are biased towards mutations characteristic of known VOCs, and therefore are not likely to provide a representative picture of all circulating lineages. Similarly, if a known lineage acquires new mutations that are not targeted by the specific SNP PCR assay in use, these will not be detected. Second, if public repositories are being used to estimate lineage proportions, and PCR pre-screening is biasing the specimens which undergo WGS and subsequent upload to repositories, the publicly available data may be more skewed. Third, PCR pre-screening might delay time to obtain genome sequence data. Furthermore, if WGS is done on a subset of samples that have already been pre-screened using the SNP PCR assays, and also used to detect and monitor other variants, the computation of expected prevalence will need to be described and adjusted for bias.

1.4.2.2. Individual-level characteristics

Some variants have phenotypic characteristics that are potentially concerning due to their ability to spread more easily from person to person, cause more severe disease, or dampen the impact of available public health and social measures (PHSMs), diagnostics, therapeutics and vaccines.

Phenotypic characteristics identifiable by clinicians and public health agencies may be used to prioritize specimens for genomic sequencing. These include specimens from:

- cases of SARS-CoV-2 infection in people who have been fully vaccinated,
- cases of SARS-CoV-2 infection in people who have been previously infected;
- cases where there is unexpected discordance between diagnostic tests, such as in clusters of individuals testing positive by rapid antigen test but negative by RT-PCR (or vice versa); characteristic and recurrent drop-out in a single gene target in a multi-target PCR assay; or where sample compartment test results are discrepant (e.g. upper versus lower respiratory tract)
- patient groups with underlying conditions that increase the likelihood of prolonged viral replication and shedding, such as immunocompromised patients (19–21)
- case clusters with unusual clinical presentations (e.g. unusually severe disease, unusual symptoms)
- case clusters suggestive of zoonotic transmission (e.g. among people working with animals susceptible to SARS-CoV-2 infection)
- cases with unexpectedly poor response to therapeutics.

Alternatively, targeting based on epidemiologic characteristics, such as travel history, particularly recent travel to an area with a high incidence of a known VOC, might be used to prioritize specimens (22).

1.4.2.3. Environmental Characteristics

Detection of variant sequences in wastewater can flag the circulation of a variant and assist in targeting further investigations and sequencing in a given geographical area (e.g. informal settlement) or setting (e.g. prison, long-term care facility, passenger ship) where randomized sequencing might be a challenge.

1.5. Metadata for genomic surveillance

All sequences should be associated with a minimum set of linked information, called metadata, which are described in WHO's Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health.

In addition to the metadata described in the above document, further variables are valuable for in-depth epidemiological analyses to characterize variants and their public health risk. Identifying these variables will likely require engagement with different stakeholders in disparate clinical and public health systems (e.g. medical records, diagnostic laboratory, vaccination services), and not all specimens are likely to have all associated metadata. However, the sharing of data publicly across systems will facilitate rapid and comprehensive evaluation of SARS-CoV-2 variants.

Table 4 below describes three tiers of metadata, with decreasing priority:

- **Highest priority**: Core metadata should always include at least the date and location of sample collection (country and state or province). Information on the place and time of specimen collection are necessary to track the spread of variants. The originating diagnostic laboratory, the laboratory conducting sequencing and the host species (human versus animal) are also minimum requirements.
- **Second priority:** The second metadata tier – which is key to triggering further investigations on characterization – is descriptive. It brings context to the genome sequence information and sequencing goals. Regardless of the sampling strategy used, the second tier of metadata includes patient characteristics (age, gender, race and ethnicity as relevant) and epidemiological characteristics (e.g. date of exposure and onset) associated with a VOC or VOI.
- **Third priority:** The third tier, which is metadata for characterization, is most helpful for analytic work to characterize the public health risk of a specific variant. Examples of variables here may include the diagnostic assay used to identify a laboratory-confirmed case, cycle threshold (Ct) value, markers of clinical severity, vaccination status, patient comorbidities, the number of secondary cases per case, travel history, association with a known outbreak or cluster or location of exposure, exposure to potentially or known infected animals, past history of SARS-CoV-2 infection and occupation as a health worker. These enhanced metadata might only be available in some settings but will greatly increase the ability to characterize the risk.

When uploading relevant metadata to public sequence data repositories, caution should be taken not to share metadata that will permit identification of individuals. It may be appropriate to share fewer data on public databases than on secure databases that are held and analysed by public health agencies.

Table 4: Recommended metadata standards to be collected for SARS-CoV-2 sequencing data

Metadata	Label	Details	Potential Analyses
Tier 1: Core metadata	Sample Identification number		
	Sample type	Examples: “sputum”, “blood”, “serum”, “saliva”, “stool”, “nasopharyngeal swab”, “wastewater”	
	Sample collection date		Introduction and evolutionary rates
	Country of collection		Introduction and transmission routes, using BEAST (Bayesian Evolutionary Analysis Sampling Tree)
	State/province of collection		
	Originating diagnostic lab	Where the clinical specimen or virus isolate was first obtained	
	Sequence submitting lab	Where sequence data have been generated	Sequencing capacity assessment
	Sampling method	Part of routine surveillance or focused sampling, representative or targeted sampling	
	Host	e.g. human, animal (specifics), environment, unknown	Transmission routes
Tier 2: Descriptive metadata	Age		Risk factors
	Sex	e.g. male, female, other, unknown	Risk factors
	Race and/or ethnicity*		Risk factors
	Heath worker status	e.g. yes, no, unknown. See HW definition in surveillance protocol for Health Workers	Transmission routes, risk factors
	Travel History	Location(s) and timing	Introductions and transmission routes
	RT-PCR assay used (if any)		
	RT-PCR Ct value (if any)		
	Symptomatic	e.g. yes, no, unknown	Severity analysis
	Vaccination status (for humans)	Date of vaccination (dose 1 and/or dose 2, as needed), vaccine type, source of information (documented evidence such as vaccine register or vaccine card versus	Vaccine failure
			Delay between onset and sequence submission
			Severity analysis
			Severity analysis
			Severity analysis
			Severity analysis
			Reinfection risk
			Therapeutic failure
			Cluster/outbreak analysis, transmission routes
			Transmission routes
			Risk factors