VIEWPOINT: COVID-19

Tackling the pandemic with (biased) data

Data are crucial for understanding and addressing the pandemic, but there are pitfalls

By **Christina Pagel**[1] *and* **Christian A. Yates**[2]

Accurate and near-real time data about the trajectory of the COVID-19 pandemic have been crucial in informing mitigation policies. Because choosing the right mitigation policies relies on an accurate assessment of the current state of the local epidemic, the potential ramifications of misinterpreting data are serious. Each data source has inherent biases and pitfalls in interpretation. The more data sources that are interpreted in combination, the easier it is to detect genuine changes in the course of the epidemic. Recently, in many countries, this has involved disentangling the varying impact of rising, but heterogeneous, vaccination rates, relaxation of mitigations, and the emergence of new variants such as Delta.

The exact data collected, and their accuracy will vary by country. Typical data common to many countries are: numbers of tests, confirmed cases, hospital and intensive care unit (ICU) admissions/occupancy, deaths and vaccinations (*1*). Many countries additionally sequence a proportion of new positive tests to identify and track emerging variants. Some countries also now collect and publish data on infections, hospitalisations, and deaths by vaccination status (e.g., Israel, UK). Stratifying all available data by different demographic factors (e.g., age, location, measures of deprivation, ethnicity) is crucial for understanding patterns of spread, potential impact of policies and efficacy of vaccines (age, timing of breakthrough infections and prevalent variants).

It is also necessary to be aware of what data is not being collected. For instance, persistent symptoms of COVID-19 (Long Covid) were recognised as a long-term adverse outcome by the autumn of 2020. However, no simple diagnostic test has been associated with the up to 200 different symptoms (*2*). Counting Long Covid relies on a clinical diagnosis, based on a history of having had COVID-19 and a failure to fully recover, with development of some characteristic symptoms, and with no obvious alternative cause (*3*). These features make it very difficult to measure routinely and so it rarely is. As a result, Long Covid is often neglected in epidemic decision-making. Failure to account for the disease load associated with Long Covid may lead to unnecessary long-term societal health burden.

The feedback between different types of outcomes, different severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants, different mitigation policies (including vaccination) and individual risks (a combination of exposure and clinical risk) is complex and must be factored into both interpretation of data and the development of policy. Using all available data to quantify transmission is crucial to ensuring rapid and effective responses to early phases of renewed exponential growth and to evaluating how well mitigation measures are working. Relying too much on a single data source, or without disaggregating data, risks fundamentally misunderstanding the state of the epidemic.

The inherent biases and lags in data are particularly important to understand from the point of view of policy makers. Because of the natural timescales of COVID-19 disease progression (see the figure), policy changes can take several weeks to show in the data, while purely reactive policy making is likely to be ineffective. When cases are rising, increases in hospital admissions and deaths will follow. When a new variant is outcompeting existing strains, it is likely to become dominant without action to suppress. The precautionary principle suggests acting early and emphatically. Conversely, when releasing restrictions, it is vital that governments wait long enough to assess the impact before continuing with re-opening.

The most up to date indicator of the state of the epidemic is typically the number of confirmed cases, as ascertained through testing of both symptomatic individuals and those tested frequently regardless of symptoms. Symptom-based testing is likely to pick up more adults and fewer younger individuals (*4*). Infections in children are harder to pick up: children are more likely to be asymptomatic than adults, are harder to administer tests to (particularly young children), are often exposed to other viruses with similar symptoms (e.g. colds and RSV) and can present with symptoms that are atypical in adults (.e.g abdominal pain or nausea). Children under 12 are not routinely offered the vaccination. Their mixing in schools provides ongoing opportunities for the virus to circulate, so it will be important for countries to keep track of infections in children as accurately as possible. Other testing biases include test accessibility, reporting lags, and the ability to act lawfully upon receiving a positive result.

Substantial changes in the number of people seeking tests may further confound case figures (*5*). Case positivity rates may provide a more accurate reflection of the state of the epidemic (*6*) but are themselves dependent on the mix of symptomatic and asymptomatic people being tested.

SARS-CoV-2 variants have been an important driver of local epidemics in 2021. The four main SARS-CoV-2 variants of concern, to date, have been B.1.1.7 (Alpha), B.1.351 (Beta), P.1 (Gamma) and B.1.617.2 (Delta). Some have been more transmissible (Alpha), some have substantial resistance to previous infection or vaccines (Beta), and some have elements of both (Gamma and Delta) (*7*). Currently, the high transmissibility of Delta combined with some immune evasion has made it the world's dominant variant. Determining which variants pose a significant threat is difficult and takes time, particularly where many variants co-circulate. This is especially true for situations where a dominant variant is declining, and a new one growing. While the declining variant remains dominant, its decrease masks increases in the new variant, as case numbers remain unchanged or fall overall. Only when a new variant becomes dominant does its growth become apparent in aggregated case data, by which time it is, by definition, too late to contain its spread. Exactly this dynamic has been observed across the world with Delta over the second and third quarters of 2021.

With multiple variants circulating, there are, effectively, multiple epidemics occurring in parallel and they must be tracked separately. This typically requires the availability of sequencing data, which is unfortunately limited in most countries. Sequencing takes time and so it is typically a few weeks out of date. These lags, and the uncertainty in sampling can lead to hesitancy in acting. The rapid path to dominance of the Delta variant in the UK highlights the need for action when a rapidly growing variant represents only a few percent (or less) of overall case load.

[1]University College London, London, UK. [2]University of Bath, Bath, UK. Email: c.pagel@ucl.ac.uk, c.yates@bath.ac.uk

Hospital admissions or occupancy data do not have the biases associated with testing behaviours and provide unequivocal evidence of widespread transmission, its geography, and demographics. However, hospital admissions lag infections more than reported cases, rendering these data less useful for proactive decision making. Hospital data are also biased towards older people who are more likely to suffer severe COVID-19, and now, unvaccinated populations. Intensive care occupancy data show a younger age profile than admissions since younger patients have a better chance of benefitting from the invasive treatment procedures (*8*).

Deaths are the most lagged indicator – typically occurring 3 or more weeks post infection and with an additional lag in registration and reporting. Death data should never be used to inform real-time policy decisions. Instead, deaths are an unambiguous eventual measure of the success of a country's epidemic strategy and implementation. The age distribution of those who eventually die from COVID-19 is different again from other metrics of the epidemic – skewed furthest towards older age groups (*9*). Those with clinical risk factors (immunodeficiency, obesity, existing lung conditions etc), high exposure (healthcare workers, low-income workers) and the unvaccinated are over-represented in COVID-19 deaths.

In countries with high vaccination rates, it is clear that vaccination has had a significant impact - reducing COVID-19 cases, hospitalisations and deaths. However, when looking at the raw numbers in highly vaccinated populations it can be the case that more fully vaccinated people are dying of COVID-19 than unvaccinated. If these raw statistics are misinterpreted, or worse deliberately misused, they can damage vaccine confidence. In reality, more vaccinated people may die than unvaccinated because such a high proportion of people are vaccinated (*10*). This does not mean vaccines are not effective at preventing death. Looking at the rates of death in vaccinated and unvaccinated individuals separately demonstrates that vaccines provide significant protection against severe disease and death. This example illustrates how important it is to curate and manage the way in which data are presented in the midst of an epidemic.

Each country has established its own vaccination priority lists and dosing schedules in order to best achieve its goals (*11*, *12*). Each of these strategies will manifest differently in the data. Additionally, many countries are using multiple vaccines in tandem and employing them differently for different demographics. Some countries are vaccinating adolescents and others are not or not offering them the full approved dose. Most vaccines require two doses, spaced between 3 and 12 weeks apart, except for the Johnson & Johnson single dose vaccine. This matters, particularly as different variants spread, because different vaccines have different effectiveness after 1 and 2 doses, different timelines to full effectiveness, and different effectiveness against variants [for instance, mRNA vaccine-mediated immunity is less affected by the Beta variant than immunity from vaccines based on adenoviruses (*13*)].

Data published on the vaccination delivery itself must thus go beyond the raw numbers of people vaccinated. Vaccine uptake must be reported by whether fully or partially (1-dose in a 2-dose regimen) vaccinated and using the whole population as a denominator. It is vital to disaggregate vaccine data by age, gender, and ethnicity as well as location so that it is possible, for example, to understand the impact of deprivation on vaccine coverage or vaccine hesitancy in particular demographics. When interpreting vaccination data, it is important to remember there is also a lag between delivery and the build-up of immunity.

Data on re-infection and post-vaccination (breakthrough) infection are also important in order to determine the relative benefits of infection-mediated and vaccine-mediated immunity and the length of protection offered. Studies that show those who were immunized earlier were catching COVID-19 with higher rates than those vaccinated more recently may suggest waning vaccine protection (*14*). Such studies have already prompted vaccine booster programmes in some countries. However, any study suggesting waning immunity must be extremely careful to ensure the 'early' and 'recent' subgroups are properly controlled. Differences in prior exposure, affluence, education-level, age, and other demographic factors between these cohorts may be enough to explain the disparities in SARS-CoV-2 infection rates even in the absence of waning immunity. Waning immunity must also be reported separately for different outcomes: for instance there might be waning in terms of preventing symptomatic infection but far less or none in preventing death (*15*). In addition, there are ethical concerns surrounding mass booster programmes in rich countries whilst many poorer countries have been unable procure vaccines to protect the majority of their populations.

Moving into the vaccination era, reported cases, hospitalisations and deaths should also be disaggregated by vaccination status (and by which vaccine), which will be easier in countries where national linked datasets exist. Whilst many sources of data are already available, this finer-grained information would help understanding of emerging issues including breakthrough infection, reinfection, new variants, and waning immunity. Additionally, incorporating Long Covid into routine reporting and policy making is crucial. Consistent diagnostic criteria and well-controlled studies will be vital to this effort. These elusive data will be of crucial importance to navigate our way successfully out of the epidemic.

REFERENCES AND NOTES

1. M Roser *et al.,* Our World Data (2021) https://bit.ly/3kepLgw
2. H.E. Davis *et al.*, E. Clin. Med. 2021; 38 101019
3. M. Sivan *et al.*, BMJ 2020; 371: m4938.
4. S.M. Moghadas *et al.*, Proc. Natl. Acad. Sci. 117 17513 (2020).
5. J. Wise *BMJ*. 370 m3678 (2020).
6. M. Hartman, COVID-19 Testing: Understanding the "Percent Positive" (2021) https://bit.ly/3CeN8wl.
7. C.E. Gómez *et al., Vaccines* 9 243 (2021).
8. A.B. Docherty *et al.*, BMJ. 369 m1985 (2020).
9. ONS, Deaths registered weekly in England and Wales by age and sex: covid-19 (2021) https://bit.ly/3Ci2obS.
10. C Yates Significant proportions of people admitted to hospital, or dying from covid-19 in England are vaccinated—this doesn't mean the vaccines don't work (2021) https://bit.ly/3kfqIoH
11. CDC COVID-19 Vaccine Information for Specif2c Groups (2021) https://bit.ly/39aijwp
12. JCVI Priority groups for coronavirus (COVID-19) vaccination: advice from the JCVI (2020) https://bit.ly/2VIhfwC
13. J.P. Moore. *JAMA* 325 1251 (2021).
14. Y. Goldberg *et al.* Waning immunity of the BNT162b2 vaccine: A nationwide study from Israel (2021) https://bit.ly/3kgDaV9
15. PHE, Duration of Protection of COVID-19 vaccines against clinical disease (2021) https://bit.ly/3CCoVAq

ACKNOWLEDGMENTS

CP and CY are both members of Independent SAGE: https://www.independentsage.org/

DOI

PHOTO CREDIT GOES HERE

COVID-19 infection progression

An approximate timeline from infection with COVID-19 to death including the times at which these figures are expected to show up in the different data sources. Although death is shown as an end point, it should be noted that most people infected with COVID-19 will survive.