What can – and can't – we learn from the official Covid-19 data?

Reflections on data and dashboards

Tom Moultrie (1987W) - Professor of Demography, Centre for Actuarial Research (CARe), University of Cape Town.

The spread of Covid-19, the disease caused by the SARS-COV-2 virus, represents an unprecedented global health crisis. Despite nay-sayers' contention that the case fatality rate is not that high, the sheer volume of people infected (where emerging evidence suggests that 70 per cent of those infected will be asymptomatic but still contagious) means that even a small proportion of infections requiring advanced health care has the potential to severely strain or even overwhelm the health system, as observed in Italy, the UK, and New York City.

The last roughly equivalent public health emergency was the Spanish Flu epidemic of just over a century ago. But then, much less data were collected and they did not flow with the ease or volume they do now. The first major global public health crisis of the digital age has spawned a huge number of online 'dashboards' – aggregations of information scraped from other online sources or official / government communiques. Some of these dashboards are global in their ambition and scope; others are set up to present or re-present the data from a given country or region.

Anecdotal evidence suggests that most of us have a few favourite such dashboards that are looked at, contemplated, and read for signs with all the avidity of soothsayers reading entrails or tealeaves.

However, these dashboards – often presented with shiny data visualisations – offer little prospect for meaningful insight into the current condition of the outbreak. There are several reasons for this.

The data from official sources are generally thin and largely uninformative. In South Africa, the Minister of Health releases on a daily basis (though not at the same time every day, as is done in some other countries) summary data showing the number of new cases identified since the previous report, classified by province; a cumulative number of tests performed; and a cumulative number of 'Covid-related' deaths. Each of these elements is problematic.

On the testing data:

- The number of new cases are not those actually testing positive since the previous report. There is a one or two-day delay between specimen collection and testing results being transmitted to the Ministry of Health. When volumes of testing are changing, this matters.
- There may be differential delays in reporting positive and negative test results, meaning that it is impossible to calculate an accurate positive test proportion from the data; and that the results do not reflect what has actually happened in the time period since the previous release of data.
- Testing regimes are changing almost constantly. As testing capacity came online, more people were able to be tested, at lower (or no) symptomatic thresholds. The positive test proportions at the start of the outbreak were higher than they are now, despite the clear evidence of continued spread of the virus.
- Even the huge increase in 'community screening and testing' is not random – those screened (by means of questioning or simple diagnostics) are referred for testing, meaning that those tested are probably more likely to be infected than the general population.
- The daily release contains the number of new cases by province; but the number of tests conducted in each province is not made public. Without a denominator (the number of tests

conducted in each province) it cannot be ascertained where the proportion of those testing positive is greatest. Presenting the number of positive cases per 100 000 population in each province implicitly assumes that testing is occurring in proportion to the population size of each province. There is no official, public, data to support this assumption.

- Emerging clinical data suggests that up to 70% of infected people may never exhibit symptoms, while still being infectious.

On the deaths:

- Limited (and it would appear, increasingly less) information is being revealed about those who have died from Covid. At times we are provided with information on the age, sex or province of residence of the decedent; at other times not.
- Yet even those deaths that are recorded do not reflect the full extent of the Covid-related deaths that are occurring. Official Covid deaths are those recorded as having come from a known infection with Covid, with a cause of death consistent with our understanding of the nature of the disease. Yet there will be other deaths, also from Covid, that are never identified as such but are instead recorded as being due to tuberculosis, stroke, or something else. And there will (as observed in Europe and the United States) be deaths that are not Covid-related but occur because the health system is strained or overwhelmed. Finally, as people die only once, an increasing number of deaths from Covid may result in deaths reported from other causes to decrease during the pandemic.
- The accounting for all these deaths is complex and will take some time to fully comprehend. At present, in South Africa – as in Europe – the Medical Research Council has implemented a system to report on all natural and unnatural deaths in near real-time. While an increase signalling an increase in natural deaths has yet to be picked up in the South African data, data from other countries show that the increase in deaths reported far exceeds those recorded as being directly due to Covid.

Despite these severe limitations, aggregator dashboards generally present these official data uncritically. They tend not to alert viewers to the very real constraints – which are not of their making – underpinning the data presented. At the very least, the producers of those dashboards should include a page setting out the constraints outlined above.

But aggregator dashboards then sometimes manipulate and analyse the official data in ways that are likely to compound this mis-information. Perhaps exponential curves are fitted to the reported number of cases (which does not allow for changes in the testing protocols, or the number of tests conducted). Perhaps the number of cases is presented relative to provincial (or worse, district) populations (demographers have little certainty in the population counts below a provincial level – migration makes extrapolation from past censuses or surveys increasingly difficult at finer levels of disaggregation). Neither of these are useful, or provide valuable insight, given the data constraints.

There is a very limited amount of sensible ways in which value can be added to the official data, and the better aggregator dashboards are deeply aware of that. More information does not necessarily imply better, or more accurate, information. In many instances, quite the reverse.

Is there a need for aggregator dashboards, then? Yes – since they can provide a mechanism to communicate official data to a broader public. However, they are of little value to the expert epidemiologists, clinicians, and health professionals at the forefront of fighting the pandemic.

That government and official agencies should be trying to present more useful and timely data is indisputable; however, doing so should not detract from their core missions. And – as is clear from the official data being released by agencies in developed countries, even they are struggling to compile and present such data.

While many of these dashboards are produced by people with incredibly impressive skills in data science and data visualisation, they reveal a very limited understanding of epidemiology on the part of the producers. Dashboards that uncritically re-present and re-analyse officially data, and who do not disclose how those re-presentations and alternative analyses are derived are contributing to public misinformation. Visitors to these dashboards should be aware of this, and treat the information provided by them with a healthy dose of circumspection.

* Tom Moultrie is professor of demography at the University of Cape Town. The views expressed are not necessarily those of his employer, or other professional and public entities with which he is associated.