Reading Pandemic Statistics

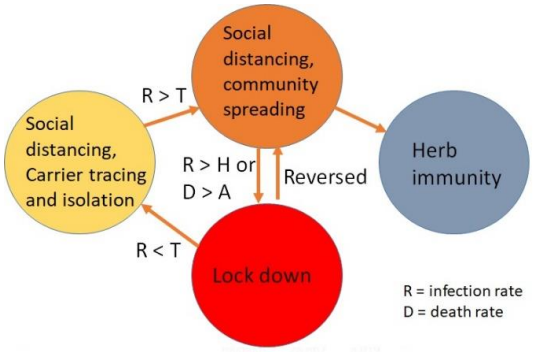
Dah Ming Chiu, April 11, 2020

It is popular for newspapers and other publishers to report the latest Covid-19 pandemic statistics using a "leaderboard" format, ranking countries (or regions) according to either cumulative *confirmed cases* or *fatality*. Locally in Hong Kong, the South China Morning Post (SCMP) shows such a table (see illustration) with every article related to the Covid-19 pandemic. Globally, the Johns Hopkins Coronavirus Dashboard, accessed 1.2 billion times daily, fashions similar content with the addition of a world map highlighting the hot spots. What can we learn from these statistics, and how do we make inferences from these statistics?

First note that these are all *cumulative* statistics, the sum of all confirmed cases and fatality from beginning to now. From these statistics, we get a very rough estimate of the scale of this pandemic, in terms of the number of people affected by it and died, so far, across the world. This let us appreciate the total damage (so far), in comparison to other calamities, such as past pandemics, wars and terrorist attacks, and other natural or manmade disasters, such as earthquakes, tsunamis and high casualty accidents. The relative numbers for different countries also let us appreciate the different scale of the problem at hand for the local governments respectively. It is tempting to make all kinds of other conclusions, but we must be very careful, since these statistics may not be accurate enough, or presented in the right way, and often we simply do not have the statistics we need, as we discuss below.

1) The number of confirmed cases is fewer than the number of infected

The cumulative number of confirmed cases is the sum of daily confirmed cases. How much the number of daily confirmed cases is less than the actual number of infected depends on how testing is done, which is linked to the disease control strategy adopted by the local government. For this reason, we take a detour to review main epidemic control strategies adopted around the world. Earlier, we wrote an article summarizing different isolation control strategies a local government can take [1]. In practice, a government tends to switch between containment controls based on the situation, as illustrated by the flowchart. As discussed in [1], in the



early stage of the epidemic, if the government can seize the opportunity, the most effective control is yellow: to track down all the carriers and suspected carriers, and isolate them, while ask the rest of the population to practice social distancing to avoid the very small number of undetected carriers. Hong Kong, I am happy to point out, is currently in this state. But when the infection rate (R) exceeds a certain threshold (T), the number of carriers becomes too numerous to be all tracked down (i.e. community spreading), then the government will likely shift to the orange control: just urge people to practice social distancing. Most western countries tried this first, e.g. Italy, Spain, France, UK and US. This situation may not be sustainable, however, for example when the infection rate begins to exhaust hospital capacity, and/or when the death rate become socially unacceptable. At this point, many governments will resort to the red control: forced lock-down, keeping everybody at home by force. Many countries enter the red state quickly, for example China, India and Philippines. This strategy is usually quite effective in stopping the outbreak, though at extremely high cost to the economy and normal daily work/life. Therefore, when it manages to cut down the infection rate sufficiently, we will return to the yellow or orange controls. China is one of the first countries making this transition, from red to orange (or yellow, depending how you look at it).

Now let us discuss how the different control strategies may affect our interpretation of the statistics. In the yellow state, most if not all the suspected and infected carriers get tested, so the reported number of confirmed cases should be very close to the true number of infected cases. In the orange and red states, the philosophy is to isolate everyone, so theoretically it is no longer necessary to discover all the carriers. Therefore, when testing resources (testing kits and manpower) are limited, maybe only the acute infected cases are discovered, leading to under-reporting. If testing capacity allows, there is still strong motivation for discovering all the carriers and try to quarantine them centrally, so there can be a big difference for different countries depending on their testing capacity and approach. Nonetheless, accurate reporting about the infection rate is crucial for social distancing by self-discipline, in the orange state. So to interpret the statistics of confirmed cases for countries in orange and red state, we need to know how aggressive they are doing testing. One indication of this is the *percentage of the population tested*. There is a website that does a great job in presenting Covid-19 data it collects – Our World in Data (OWD) [2]. They emphasize that the confirmed cases are only a subset of the total cases, and that it is important to know how testing is done, and try to provide this data (for those countries that disclose this information). They show Iceland as a good example of providing more adequate information about their testing statistics [3]. As of this writing, the statistics for some of the top countries in number of tests per 1000 (NPK) people are: Italy, South Korea and US, with 15, 9 and 7 per thousand respectively. Although many countries provide NPK statistics, some don't, for example, China. For countries that do not provide NPK statistics, especially if they are in the red state, it is hard to tell how trustworthy their number for confirmed cases is.

Finally, a brief word about death statistics, it also depends on testing as in the case of confirmed cases. Death statistics are usually more accurate, but as an indicator of current state of the outbreak, there is a bigger time lag.

2) The cumulative statistics and daily number do not tell the current situation

Another problem with the leaderboard reporting is that by using cumulative numbers, it does not necessarily reflect current situation. It so happens that many of the countries at the top of the leaderboard are higher population countries where that outbreak is peaking recently. It this were not the case, the top of the leaderboard could be the same every day, rendering the leaderboard of little information value.

One solution is to rank the countries according to daily confirmed cases instead of the cumulative number. This method emphasizes the size of the current damage. An alternative is the approach taken by OWD [2], instead using the daily number, they show the number of days it takes to double the total. By measuring the rate of change, this metric is effectively ranking countries according to how well they are slowing down the epidemic, irrespective to the size of the country.

3) For better comparison, try normalization

The public may become interested in a country (for visit or other reasons, e.g. concerned about their friends living there). The reported statistics in the leaderboard give you a comparison (across different countries), but may be misleading. At the time of writing this article, the number of deaths in the US is surpassing that of Italy. But the US population is more than 5 times of that of Italy, and the size of the country is also much larger. So as a country as a whole, the US situation is probably not as bad as in Italy. On the other hand, the New York/New Jersey area has around half of the Italy population, suffering close to half of the fatality of the US total. In that sense, it may be reasonable to conclude that the situation for the NY/NJ region is comparable to that of Italy.

4) Inferring other properties of the virus

Finally, it may be tempting to infer biomedical properties of the virus based on the statistics in the leaderboards. For example, one key property is the Case Fatality Rate (CFR), namely the number of fatality per number of infected. If you take the numbers in the table in the beginning of this article, 102564/1694247 = 6%. This is far higher than the CFR reported in scientific publications so far. Due to difficulty in sampling, this is still not conclusive. There are all kinds of other properties of interest: how does this virus affect different age groups? Different gender? People of different races? Under different environmental conditions, such as temperature? The reported statistics are just too rough and not enough to answer these and other questions.

Conclusion:

In this short article, we discuss what we can learn from the leaderboard statistics of the Covid-19 pandemic, and what we must be careful interpreting. Given its high exposure, we wish it can be more informative. For a better source of similar statistics, we recommend OWD [3].

References

[1] Isolation Control and Its Challenges, http://personal.ie.cuhk.edu.hk/~dmchiu/isolation2.pdf
[2] Our World in Data, https://ourworldindata.org/coronavirus
[3] Covid-19 in Iceland – Statistics, https://www.covid.is/data