Analysis of ECDC Covid-19 Data - 2020-03-20

Introduction

As of March 20, 2020, the Covid-19 virus has spread around the world and killed almost 7,000 people. Within a week the total number of deaths worldwide will likely exceed 10,000.

As I've been hunkered down in my house in Chicago this week following the news about the ongoing global spread of the virus, I came across the website of the European Centre for Disease Prevention and Control and discovered that they have been publishing daily updates to a dataset with numbers of new daily confirmed COVID-19 cases and deaths grouped by country and date.

I decided to dig into the data set a bit and generate some time-series plots in R to try to better understand the true severity of the virus and how it may continue to spread and impact the lives of millions or even billions of people around the world. Like many people, I am alarmed by the degree and manifestations of the virus' contagion, and I wanted to look at the data to try to better-understand the range of possible outcomes in the weeks and months ahead.

I have no medical or epidemiological training whatsoever, but I have done a fair bit of data exploration and analysis in my career. So I am going to share some of my observations and analysis of the data, but please bear in mind that this analysis is in no way rooted in any understanding of how viruses spread or any knowledge of epidemiological modeling or best-practices.

When I have some time later this week or early next week, I'll post the simple code that I used to generate the following plots on github in case anyone wants to fetch the latest data set from the ECDC and run updated plots. I may post some more notes and code for generating additional plots.

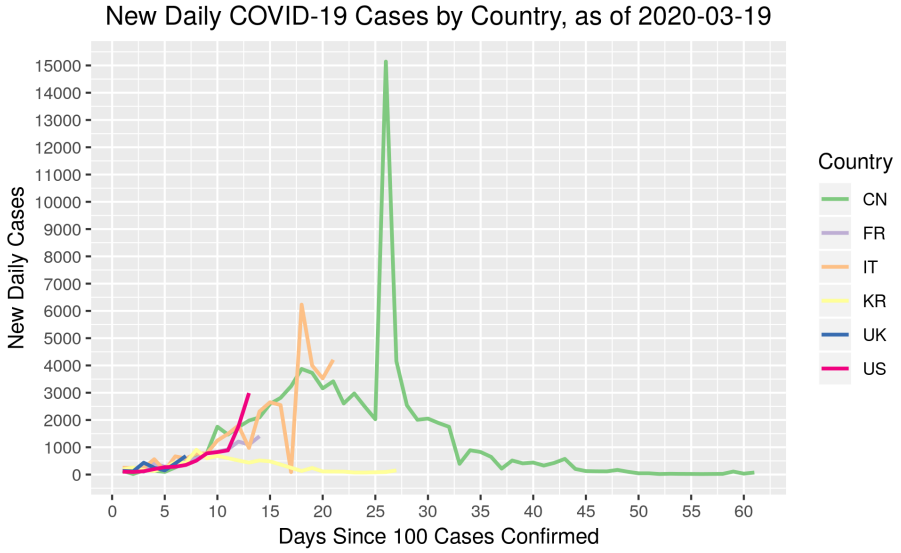
New Confirmed Cases

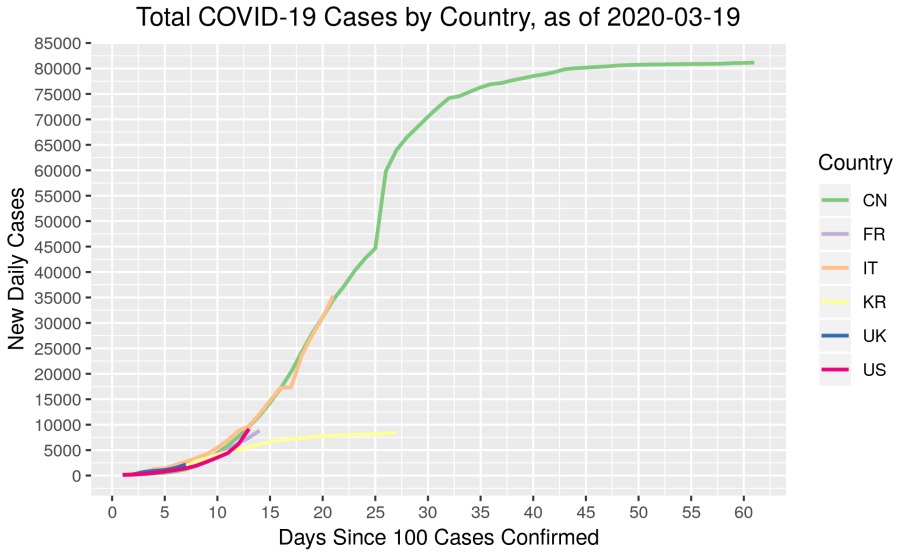
The ECDC data set includes data for daily confirmed new cases of COVID-19 as well as data on daily deaths.

I plotted time-series curves of new daily cases by country starting with the first day that confirmed new daily cases in any given country exceeded 100. This seemed like a reasonable way to compare the spread of COVID-19 across multiple countries where the virus was introduced on different dates. Some internet searching has uncovered several other academic studies that used the exact same criteria to compare the growth of infections across multiple countries (classify day 1 as the first day on which confirmed cases exceed 100).

It's clear from both the data and from news reports that the actual number of COVID-19 infections around the world is greater than the number of confirmed cases. This is likely because many healthy people – especially healthy younger

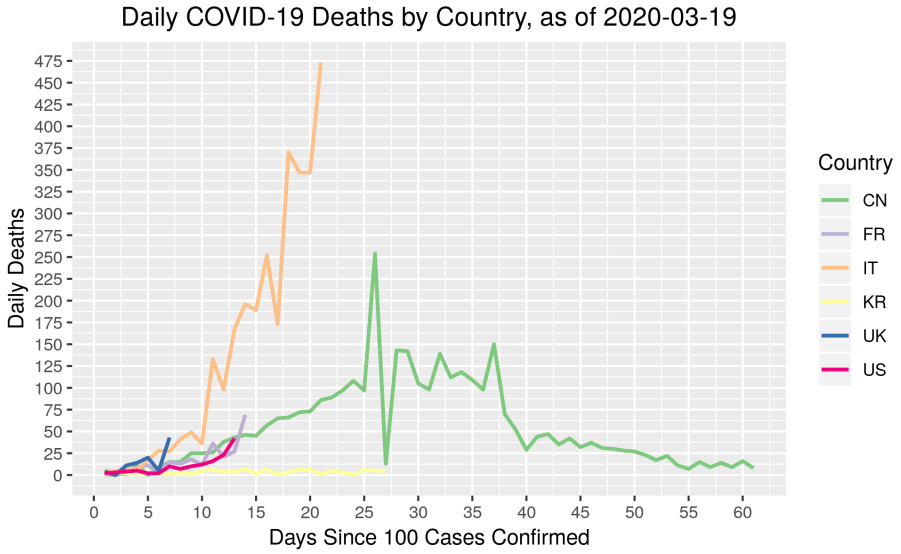
adults and children – do not develop severe symptoms when contracting the virus *and* because the availability of COVID-19 tests has been limited in many places where outbreaks have occurred. In the USA, several NBA players have tested positive for COVID-19 but have publicly stated that they have experienced no cold-like symptoms. These anecdotal cases demonstrate that it's possible that many people with no symptoms can be carrying and spreading the virus.

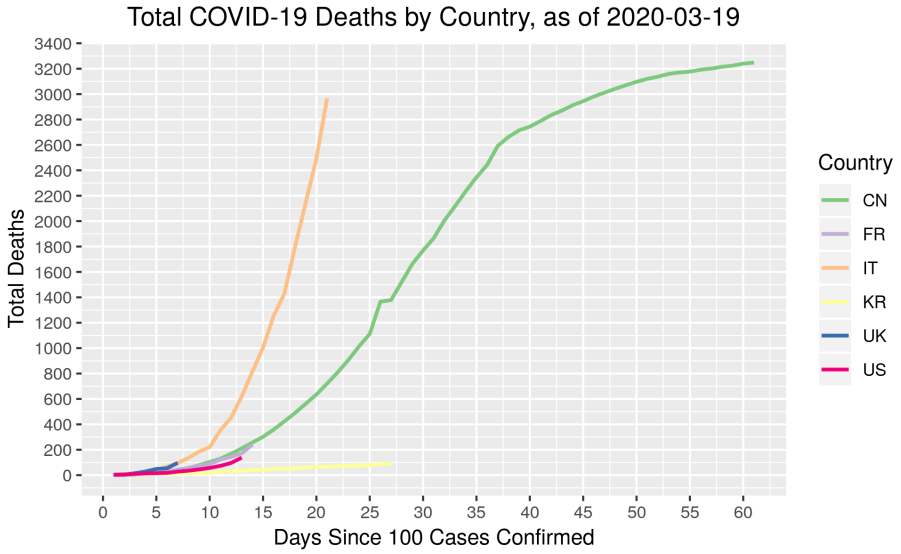




Deaths

The plots below show new deaths and the cumulative total of deaths by country by date.





Some observations:

- South Korea seems to have contained the spread of COVID-19 and reduced their case fatality rate much better than several other countries have been able to do. From what I've read, they developed testing kits and rolled

out drive-through testing sites within a couple weeks of their first cases. After several weeks of reported cases, the USA has still not manged to systematically deploy testing on the same scale even in areas with major outbreaks (e.g., New York City).

- The number of deaths in Italy has been tragically high. I read a couple articles that attributed this in part to the belief that Italy has been categorizing many deaths that occur in hospitals as COVID-19 deaths even if the patient who died had a serious pre-existing medical condition. That argument suggests that some at-risk elderly patients in hospitals whose deaths have been classified as COVID-19 deaths would have died anyway from their other conditions. Many observers have noted that Italy has the second-oldest population globally, and a major reason for the high case fatality rate is that many elderly patiIents in Italy contracted COVID-19. The demographic argument is plausible, but I suspect that the total number of cases in Italy is much greater than the number of confirmed cases and the health system has been unable to test large numbers of people due to the large scale and rapid spread of the COVID-19 outbreak in Lombardy.
- The curves for France, the UK and the USA are not significantly less steep than the curves for China and Italy. I am concerned that COVID-19 spread widely before governments mandated social distancing measures in Paris, London and New York, and it seems that there could be thousands of deaths in these cities. I saw in the news tonight that the UK government just announced today (2020-03-20) that they are shutting down all pubs, restaurants and clubs. I think it's a good move, but given that there were 43 recorded COVID-19 deaths yesterday in the UK, it appears that the virus has already spread widely. The curve in the plot of daily deaths in the UK is similar to the curve for Italy but about 14 days behind. If it holds then the UK would have thousands of deaths from COVID-19 in the next couple weeks. I hope that won't materialize.

I am hoping that people heed government orders to shelter in place and that social distancing and diligent handwashing will slow the rate of fatalities

Death Rate

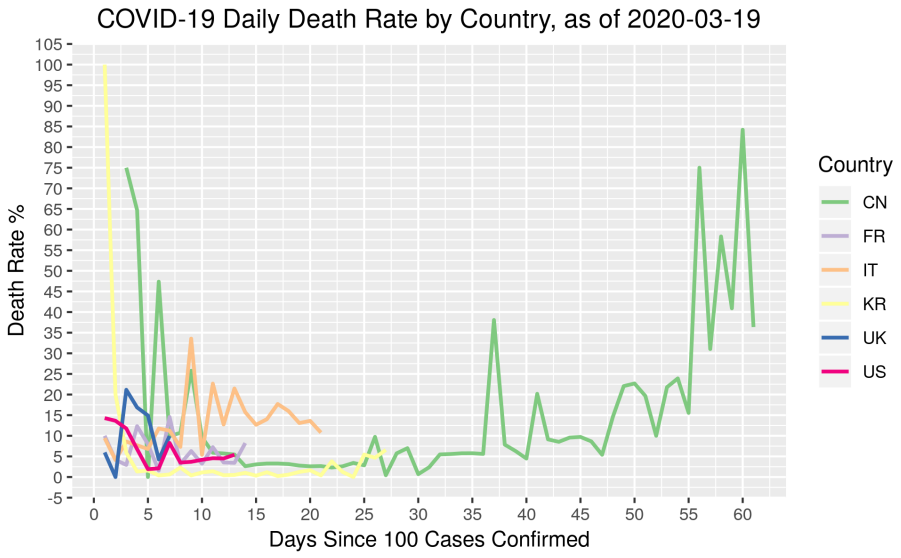


Figure 1: Death Rate

Studies by public health experts that I found on the internet suggest that the case fatality rate (CFR – the percentage of infections that ultimately result in death) of COVID is between 0.5% and 8%. Others have observed that the CFR in Italy seems to be as high as 8%, whereas the CFR in Wuhan was about 1.4%.

Using the ECDC data, I created a "death rate" plot by country, which takes the recorded number of deaths divided by confirmed cases from 4 days earlier. The number of days is somewhat arbitrary. This rate is *NOT* any sort of rigorous CFR. When determining CFR, I believe that epidemiologists consider that infections are likely underreported – especially for COVID-19 given that infected people can present with minor symptoms or no symptoms, and they develop a statistical model that adjusts for this. I haven't done that at all. I saw one study that uses a formula similar to mine to approximate the fatality rate but uses a longer lag between deaths and new confirmed cases.

My assumption is that many cases are not tested and confirmed until patients develop severe symptoms and seek medical attention. Various studies have estimated that the incubation period for COVID-19 is between 2 and 14 days (one recent study suggests that the median may be around 5 days).

With this in mind, I do not believe that the "Death Rate" in my plot is a good approximation of a CFR on an absolute basis. However, I plotted this to give me a means to compare the relationship between confirmed new cases and deaths on a relative basis across countries.

Here are some observations from my comparison: - The “death rate” in China has increased since the beginning of March. Officially, China has been reporting confirmed new daily cases in the single digits in early March. Since then, there have been news reports that cases in China grew as some Chinese nationals contracted the virus outside China and presented with symptoms and were confirmed to have the virus upon returning to China. As the official number of new cases was about 20 between March 10 and March 16, there have been between 10 and 20 recorded deaths in this time period. While news reports from China seemed to suggest that they managed to contain the virus in early March, the ongoing number of daily deaths seems to suggest that it has not been contained and that the numbers of new infections are being underreported. I’m not suggesting that this is due to some government coverup. It could simply be because many people with mild or no symptoms contracted the virus, and the virus continues to spread in communities in which not everyone is being tested. With that said, the official number of new cases jumped from 25 on March 16 to 110 on March 17. This is all to say that it appears that the virus has not been contained in China, and it is possible that ther could be a new wave of infections in the coming months in China. - The “death rate” has been much lower in South Korea than several other countries. Part of this is likely due to much higher levels of testing in South Korea than in other countries, but the overall low number of deaths in South Korea does suggest that their public health system has done a better job in managing the virus than other governments have. However, several people have still been dying each day in South Korea from COVID-19 in recent days, and it does not seem that the virus has been fully contained there either.

Final Thoughts

In the past week, several local and national governments in Europe and the USA have shut down restaurants and bars, banned public gatherins, and even issued orders asking people to distance themselves from others and shelter in place.

It’s likely too early to tell whether these orders came early enough to slow transmission to the levels of South Korea and Japan or if prior widespread community transmission has already spread the virus to so many people in parts of France, the UK and the USA that we’ll see a situation like Italy’s develop in more countries. The initial curves of daily deaths in these other countries are not encouraging, but the severe government response could still make a difference.

When I get some time in the next week, I’ll post my code to generate these plots on github, and I’ll generate updated plots from the latest ECDC data on the github page.