COVID Cases and Death Correlation

by Adrian C. Keister, Ph.D.

2020/11/12

It has been argued[1] that COVID death rates lag COVID case rates by 2-8 weeks. The authors present a data visualization claiming to make clear the lag in time between cases and deaths. However, the data visualization does not do so. The visualization shows that cases rose for a while, and then deaths also started to rise somewhat. Unfortunately, the data only goes out to August 6, 2020, and does not take into account more recent death rates, which are much lower and do not fit the narrative. In the footnoted article above, it should be noted that the authors cite a WHO paper, which mentions this finding on page 14. The WHO paper says, "Data on the **progression of disease** is available from a limited number of reported hospitalized cases (Figure 5)." - p. 13, emphasis original.[2] Hence the WHO paper is reasoning from actual fatalities.

This paper seeks to investigate the claim, to determine its strength. On its face, this might seem plausible, since, reasoning as above, anecdotally, the median time from the onset of symptoms to death is in that time frame. However, reasoning from single cases to the aggregate commits the fallacy of composition and is not borne out by the data. It is, in effect, a highly biased sample, based solely on fatalities, whereas cases involve many recoveries.

To show this lack of correlation, I used data from the covidtracking project,[3] as it seems less biased than other sources. Indeed, one great hindrance to doing anything with COVID data is the highly politicized nature of it. I downloaded the all-state data and performed the following operations:

1. Eliminated all columns but date, state, positive, and deathConfirmed. The positive and deathConfirmed columns correspond to cases and deaths, respectively. It is important to note that these are cumulative numbers.

2. As reporting is highly irregular on the weekends, I smoothed out the cumulative numbers by interpolating over the weekend: Saturday's number is 2/3 Friday's number plus 1/3 Monday's number, and Sunday is similar but weighted more towards Monday.

3. Then I took a discrete difference in the cases and deaths to produce daily rates.

4. Finally, I used a series of correlations to determine which correlation was the strongest, and where it occurred. The idea here is to shift one time series past the other one day at a time, compute the correlation at each step, and find the largest such correlation. This should correspond to the lag time, if there is a reasonably strong correlation.

---

[1]*Visualizing the lagged connection between COVID-19 cases and deaths in the United States: An animation using per capita state-level data (January 22, 2020 – July 8, 2020)*, by Christian C. Testa, Nancy Krieger, Jarvis T. Chen, and William P. Hanage. HCPDS Working Paper Volume 19, Number 4.

[2]*Report of the WHO-China Join Mission on Coronavirus Disease 2019 (COVID-19)*.

[3]https://covidtracking.com