COVID-19 DATA COLLECTION: GARBAGE IN, GARBAGE OUT

HARVEY J. STEIN

Abstract. I analyze anomalies in the data collected by the major COVID-19 data repositories. These anomalies are due to how the data is aggregated. I find that the method used tends to understate the rise in the infection rate, overstate the fall after the peak, and generate spurious peaks and drops. As a result, relying on this data can lead to delaying taking action during the rise up to a peak and delaying normalization after the fall.

I make recommendations for changing the aggregation methodology to correct for these problems. Moreover, rather than aggregating the data at all, making the raw data available would both improve the situation as well as make more accurate and detailed analysis possible.

Contents

1. Introduction

COVID-19 data is collected, aggregated, and published by a variety of agencies. The data itself comes from a variety of sources, such as hospitals and testing facilities. The sources generally do not report cases on a daily basis due to various delays involved. Some sources might report weekly, and others may report monthly. As a result, the reported totals for a given date will be understated until reports from all sources arrive at the aggregating agency, which can take as much as a month's time.

My 4/21 analysis of the NYC COVID-19 data (COVID-19 NYC Stats – Not What They Seem), shows a graph demonstrating that this missing data can have a major impact on the reported incident counts. Figure 1 contains a similar graph for NYC's data around the first wave's peak.

Comparing the counts for a given date as they are updated over time shows that it can take a month or more before all of the data for a given date finally arrives.

Date: February 2, 2022.

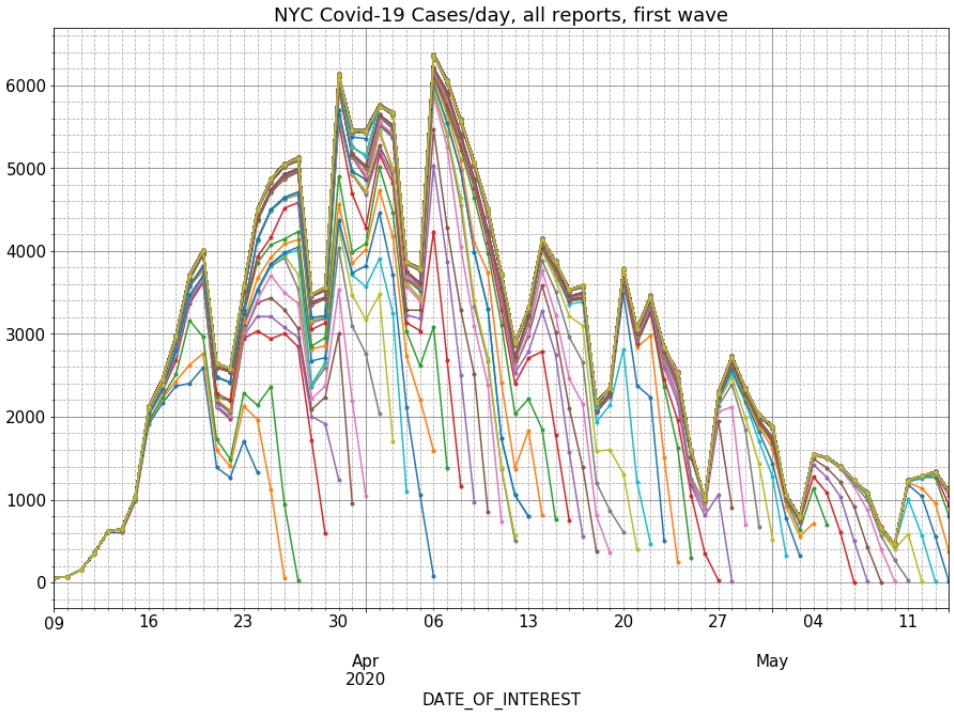

FIGURE 1. NYC COVID-19 cases per day from each daily report around the peak of the first wave, from the NYC COVID-19 data github repository. It can take a month before the totals are complete. Initial reports sometimes include as little as one fifth of the total cases.

Moreover, the final count for a given date can be five times larger than the initial count. Figure 2 has the 7 day rolling averages for the same time period. The 7 day averaging adjusts for the weekend dip that the data has, and reduces the discrepancy between the first and final counts, but still shows substantial lags in data collection.

After seeing this analysis, Jon Asmundsson, editor of *Bloomberg Markets* asked me if this holds for other regions. [Asm20]

This led me to take a look, the results of which I originally wrote up in COVID-19 data collection: Garbage In, Garbage Out. I revisit the analysis here, adding an analysis of the Omicron data.

2. Garbage in

My first stop was the Our World in Data github repository. I forked the repository, imported my analysis code, extracted the historical reports and graphed the histories. The results are in Figure 3.

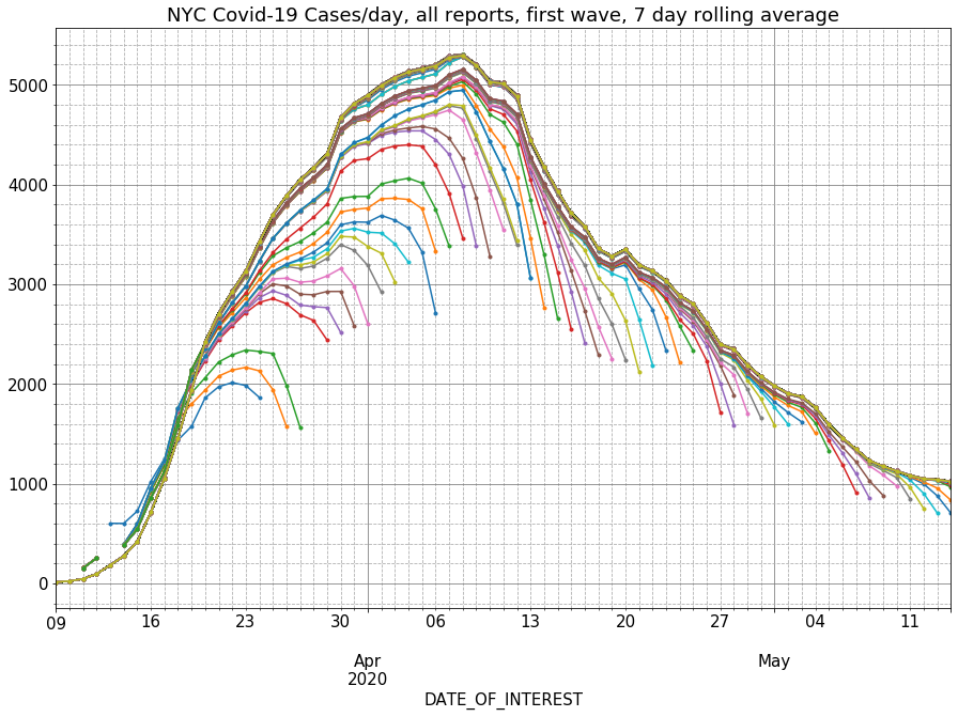

FIGURE 2. 7-day rolling average of NYC COVID-19 cases per day from each daily report, from the NYC COVID-19 data github repository. The missing data has less impact on the rolling average, but their effect is still substantial, especially near the peak.

I was surprised to find that there are no updates to past records! How can it be that the total cases for a given day are completely known the following day and never need to be updated? Given what we know about the NYC data, and given that the NYC data is a large part of the USA data, it can't possibly be the case that on a given day they know exactly the number of cases the day before. Something else must be going on.

So I entered an issue on the OWID COVID-19 github repository. I asked for clarification of how they collect and publish the data. Edouard Mathieu, Data Manager at OWID, responded:

> For confirmed cases and deaths, our data comes from the European Centre for Disease Prevention and Control (ECDC). We discuss how and when the ECDC collects and publishes this data here.
>
> Importantly, the ECDC follows a general rule of not changing past values in its data. If cases/deaths are reported with a lag – a general lag, as you described, or occasional 'blocks' of new data–these new cases/deaths will be added **on the date that the country reported them to the ECDC.** [Mat20]

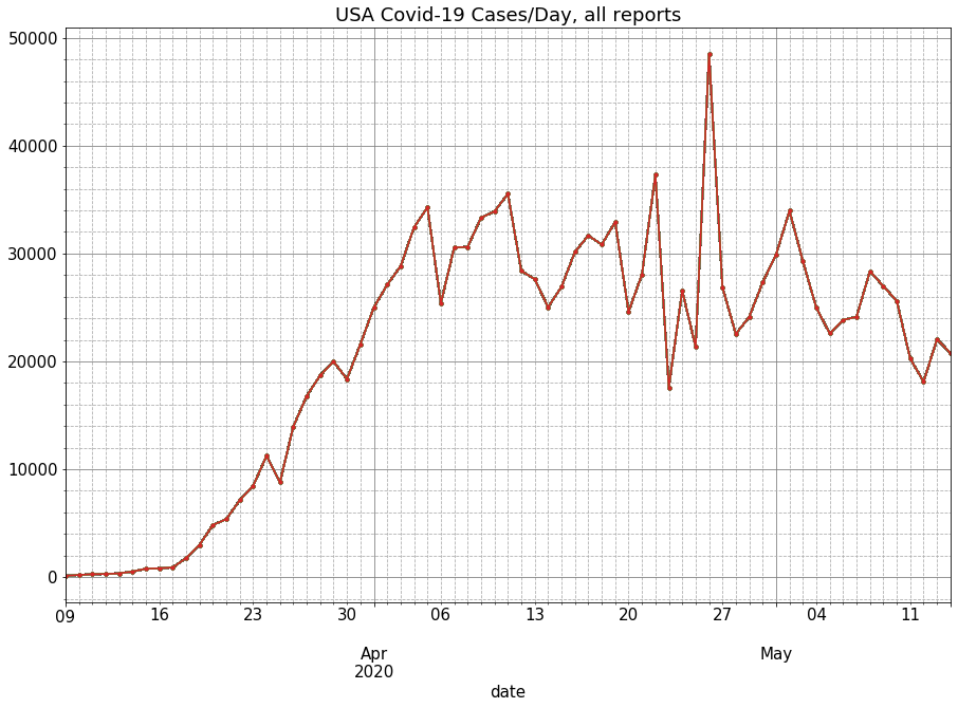

FIGURE 3. USA COVID-19 cases per day from each daily report, from the OWID COVID-19 data github repository. Newer reports never restate older totals.

So, OWID gets their data from the ECDC – The European Center for Disease Prevention and Control, and the ECDC doesn't collect data by incident date, it collects the data by the date on which it receives the reports.

Investigating other data hubs, I discovered that it's not just the ECDC. The Johns Hopkins University COVID-19 repository, and the New York Times COVID-19 repository also record instances by the report receipt date instead of by the incident date. So these databases, the major sources of data that people use for modeling, for planning disease responses, and for reporting, are collecting the data by reporting date instead of by incident date.

I followed up with Edouard Mathieu. I asked him if he knew of any rationale for why the data was being collected this way. His impression was that governments try to give the most accurate view and record data based on incident date, updating history as needed. On the other hand, aggregators like WHO, ECDC and JHU are more concerned with ease of aggregation and stability of reported numbers, so they instead record data based on the reporting date. [Mat20]

I also contacted Lauren Gardner, Associate Professor, Department of Civil and Systems Engineering, Co-Director of Center for Systems Science and Engineering (CSSE), Johns Hopkins University. Professor Gardner and her team are responsible for the Johns Hopkins COVID-19 Dashboard. She agreed that there are issues with

using reporting dates rather than incident dates, but unfortunately, that's often all that's available. [Gar20]

3. Garbage out?

What's the big deal? Counting by reporting date instead of incident essentially takes some percentage of the actual data and moves it later in time. One would expect this to flatten and shift the curve. It would make the infection rate appear lower before the peak, make the peak appear later, and make the infection rate appear to drop off more slowly after the peak. Moreover, since sites will report a number of days together, it also makes the data jumpier and thus harder to analyze.

The problem is that scientists are using these numbers to model the disease, the government is using these numbers to plan how to address the risks, and the media is reporting about the numbers. So it potentially reduces the accuracy of the models, interferes with planning and leads to hysterical media reports about fictitious rising and falling of death counts.

How big is the effect, really? To start, I calculated some simulated results. I consider the true counts to follow a normal distribution, and compare them to what happens if 30% or 50% of the data comes in late (Figure 4). It can have a substantial impact.

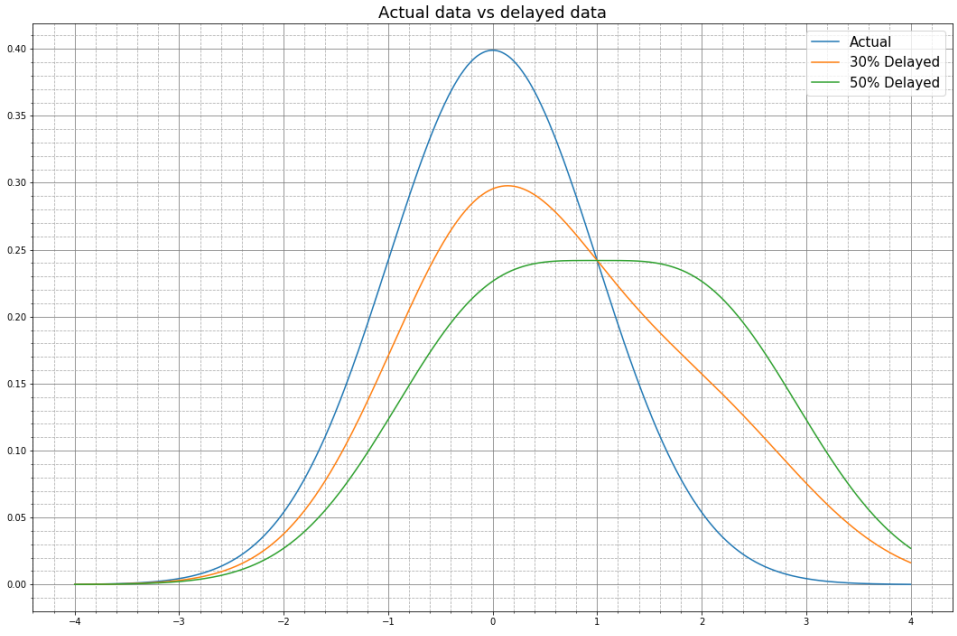

Figure 4. Impact of data being delayed. If some of the data is reported late, then it is likely to understate the rise and delay the apperance of the fall.

We can go further. The NYC data has both the reporting dates and the incident dates, so I backed out what it would look like if it was recorded by report date

instead of by incident date. Figures 5 and 6 compare the report date results to the incident date results for the first peak and for the Omicron peak.

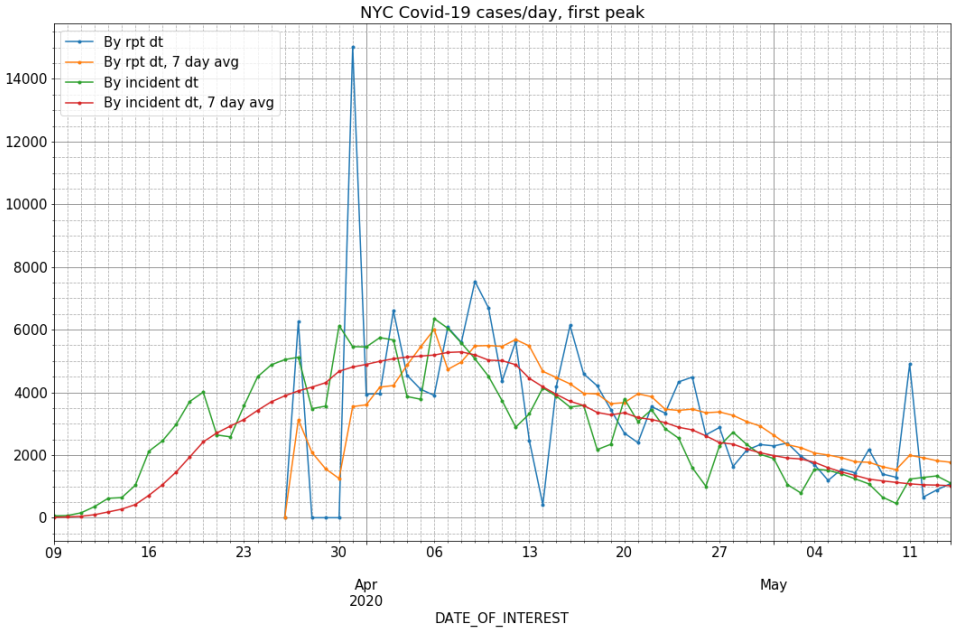

FIGURE 5. NYC cases per day, first peak, by reporting date vs by incident date. The reporting date data is very noisy and plagued by spurious peaks and drops, even after being averaged on a 7 day basis.

As you can see, the reporting date data is far noisier; so much so that the 7 day cycle I documented in Covid-19 NYC Stats - A Ray of Hope is obscured and the 7 day rolling window still shows substantial noise. For example, the report date based data exhibits a spike on May 11th of about 5,000 cases, far higher than the incident based data shows. The incident based data shows a slight restatement of the data going back to the third week of March.

Presumably NYC received a report around mid May from a particular site, and that report relayed daily infections back through March that hadn't yet been recorded. Report date based data then records this as a huge spike which never actually occurred. Because of these underlying data collection mechanisms, report date based data collection often give a substantially distorted view from day to day.

Another case in point is that the noise yields a false peak prior to the actual peak.

Even after smoothing with a 7 day rolling window, the report date data is overstating the post-peak number of cases by a substantial amount, sometimes by over 50%.

Surprisingly, the growth rate to the peak is higher rather than lower. This is presumably because reporting delays were greater when the data started being collected, leading to batches of reports coming in together at a later date.

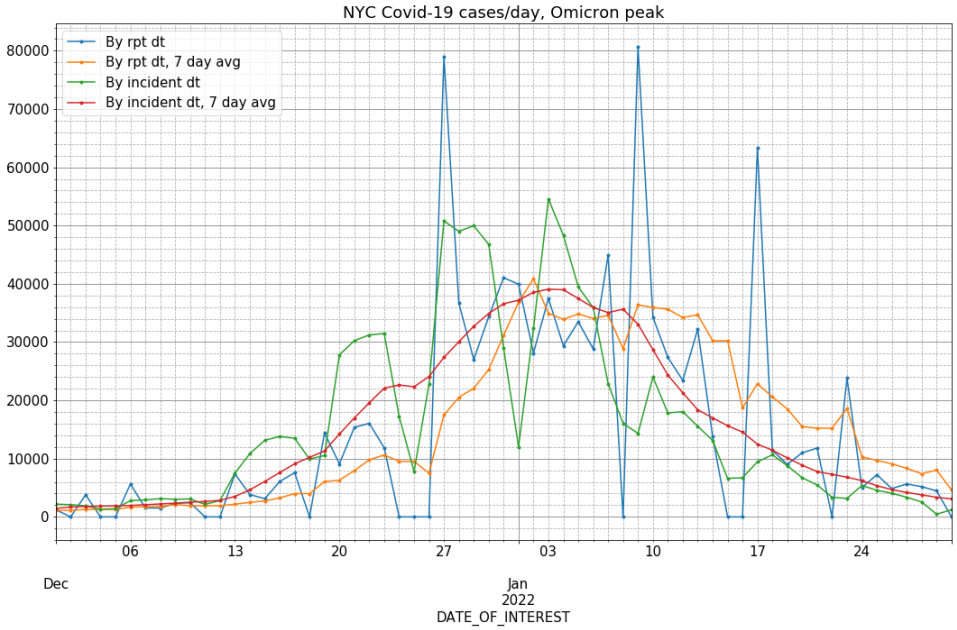

FIGURE 6. NYC cases per day, Omicron peak, by reporting date vs by incident date. The reporting date data is as poorly behaved and misleading as for the first peak.

The Omicron peak shows similar results. The counts by reporting date show exaggerated spikes that are two to six times larger than the actual counts for those dates, and the 7 day rolling average is noisy, underestimates infection rates before the peak, and overestimates them after the peak.

4. Conclusions

It is great that during the COVID-19 pandemic, organizations like OWID, the WHO, and the ECDC, major news outlets like The New York Times, and major universities, like Johns Hopkins University, are all collecting and aggregating data on COVID-19 cases and deaths and making this data publicly available. Without such aggregation, it would be very difficult to globally understand, analyze, and respond appropriately to the pandemic.

On the other hand, it's unfortunate that they collect the data in a way that obscures the current state of the disease and makes analysis more difficult than it need be. It causes the data to be unnecessarily noisy, creates false peaks, understates rates leading up to peaks, and overstates them afterwards. It's disturbing that news agencies are reporting on these numbers as if they actually occurred on the reporting date, and governments are taking action based on the same misconception.

One might be surprised that epidemiologists who make a career out of analyzing epidemics and pandemics would record the data in such a fashion. But, such work

tends to be on a longer time scale and it's only in the current pandemic that we needed accurate, real-time data.

I hope that data collection agencies will take it upon themselves to correct this and begin collecting and aggregating the data on an incident date basis instead of a reporting date basis. This would be a significant undertaking, but would better enable us to analyze the data and respond appropriately, if not to this pandemic, then to the next one. Even better would be to release the data from the raw reports. Then no guesswork would be required to account for reporting delays and analyses could be greatly enriched.

References

[Asm20] Jon Asmundsson. *Email exchange discussing dataset date recording approaches*. Private communication. Bloomberg, May 2020.

[Gar20] Lauren Gardner. *Email exchange discussing dataset date recording approaches*. Private communication. Johns Hopkins University, May 2020.

[Mat20] Edouard Mathieu. *Email exchange discussing dataset date recording approaches*. Private communication. Our World In data, May 2020.

[OWI20] OWID. *Data on COVID-19 (coronavirus) confirmed cases, deaths, and tests, All countries, Updated daily by Our World in Data*. 2020. URL: https://github.com/owid/covid-19-data.

[Ste20a] Harvey Stein. *Analysis of NYC COVID-19 infection rate*. Fork of https://github.com/nychealth/coronavirus-data. 2020. URL: https://github.com/hjstein/coronavirus-data.

[Ste20b] Harvey Stein. *Analysis of Our World of Data's COVID-19 dataset*. Fork of https://github.com/owid/covid-19-data. 2020. URL: https://github.com/hjstein/covid-19-data.

[Ste20c] Harvey Stein. *COVID-19 Data Collection – Garbage In, Garbage Out*. Blog, Harvey J. Stein, Essays and commentary from a member of the quantitative community. Apr. 2020. URL: http://hjstein.blogspot.com/2020/05/covid-19-data-collection-garbage-in_33.html.

[Ste20d] Harvey Stein. *COVID-19 NYC Stats – A Ray Of Hope*. Blog, Harvey J. Stein, Essays and commentary from a member of the quantitative community. Apr. 2020. URL: https://hjstein.blogspot.com/2020/04/covid-19-nyc-stats-ray-of-hope.html.

[Ste20e] Harvey Stein. *COVID-19 NYC Stats – Not What They Seem*. Blog, Harvey J. Stein, Essays and commentary from a member of the quantitative community. Apr. 2020. URL: https://hjstein.blogspot.com/2020/04/covid-19-nyc-stats-not-what-they-seem.html.

[Joh20] Johns Hopkins University. *Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE*. 2020. URL: https://github.com/CSSEGISandData/COVID-19.

[New20a] New York City. *NYC Health Coronavirus Data*. 2020. URL: https://github.com/nychealth/coronavirus-data.

[New20b] New York Times. *An ongoing repository of data on coronavirus cases and deaths in the U.S.* 2020. URL: https://github.com/nytimes/covid-19-data.

Email address: hjstein@bloomberg.net