COVID-19 DATA COLLECTION: GARBAGE IN, GARBAGE OUT

HARVEY J. STEIN

Abstract. COVID-19 data discussion.

Contents

1. Introduction

If you recall my 4/21 analysis of the NYC COVID-19 data (COVID-19 NYC Stats – Not What They Seem), you'd remember a graph showing the extent to which missing data has a major impact on the reported incident counts. Figure 1 has an updated version.

Date: May 14, 2020.

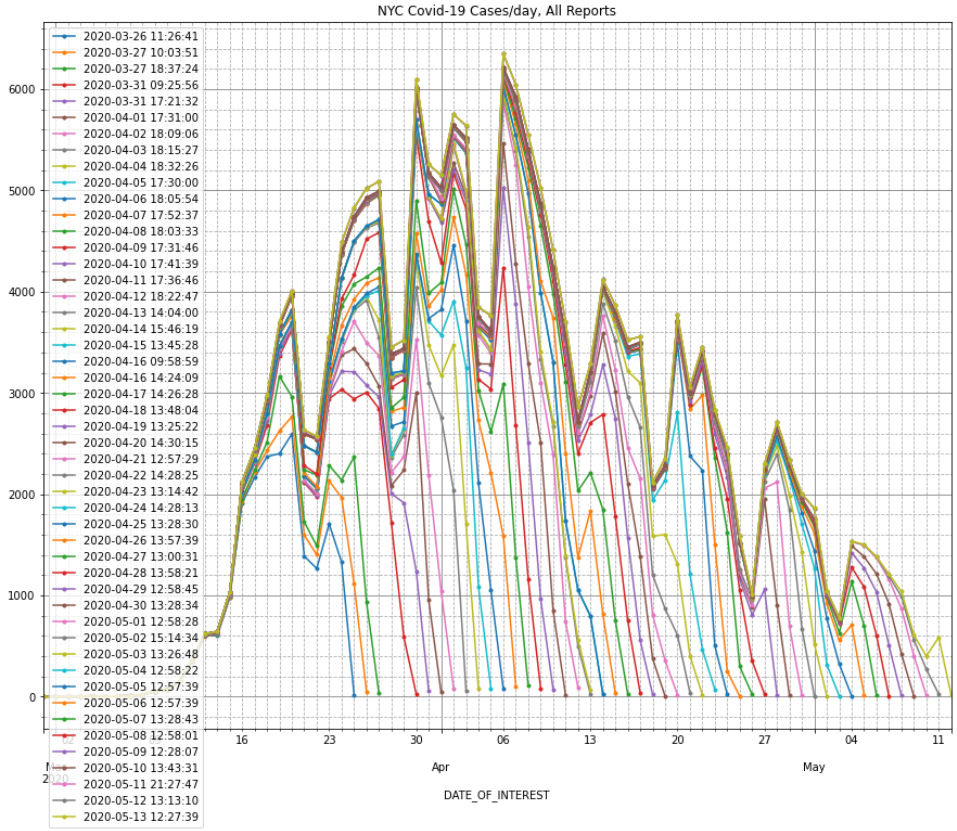

FIGURE 1. NYC COVID-19 cases per day from each daily report. Data from the NYC COVID-19 data github repository.

As you can see, the peak hasn't moved from April 7th, but we're still getting data for dates as far back as March! Figure 2 has the updated 7 day average report, which shows more clearly that three days ago additional incidents were reported for every date going back as far as March 25th.

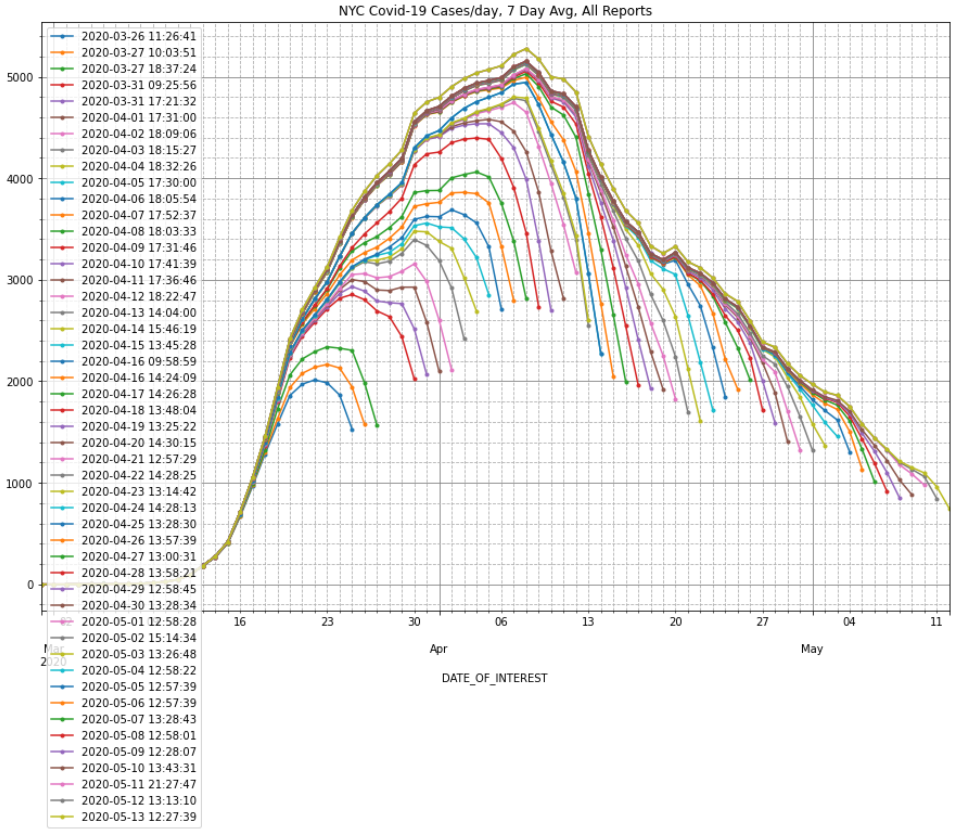

FIGURE 2. 7-day rolling average of NYC COVID-19 cases per day from each daily report. Data from the NYC COVID-19 data github repository.

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

This led me to try.

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 history. The results are in figure 3.

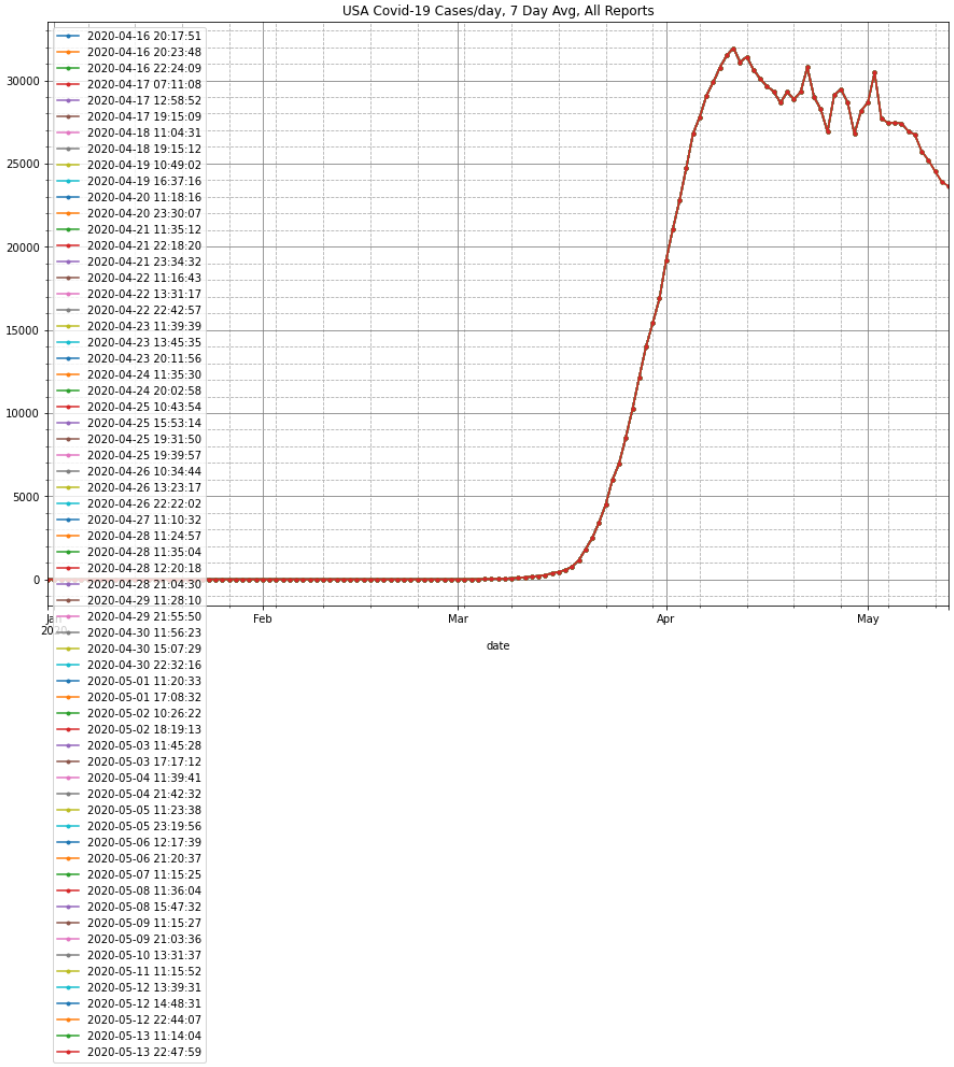

FIGURE 3. USA COVID-19 cases per day from each daily report. Data from the OWID COVID-19 data github repository.

It's nice that the 7 day average is dropping, but where are the new reports? There are no updates – no missing data! How could it be that the data the USA cases counts for yesterday that they receive today are complete? Given what we know about the NYC data, and given that the NYC data is 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 odd must be going on.

So I entered an issue on the OWID COVID-19 github repository. I asked how they generate 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.**

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.

Further research showed 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 report receipt date instead of by incident date. And these are the major sources of data that people use for modeling, for planning disease responses and for reporting.

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 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.[Gar20]

3. Garbage out?

What's the big deal? Counting by report date instead of incident essentially takes some percentage of the actual data and moves it later in time. One would expect this to flatten the curve. As a result, it should makes 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 reduces the accuracy of the models, interferes with planning and leads to hysterical media reports about irrelevant rising and falling of death counts.

How big is the effect, really? We can compare where we have the report dates along with the incident dates. I did this for the NYC data. I backing out what it would look like if it was recorded by report date instead of incident date. Figures 4 and 5 give the results.

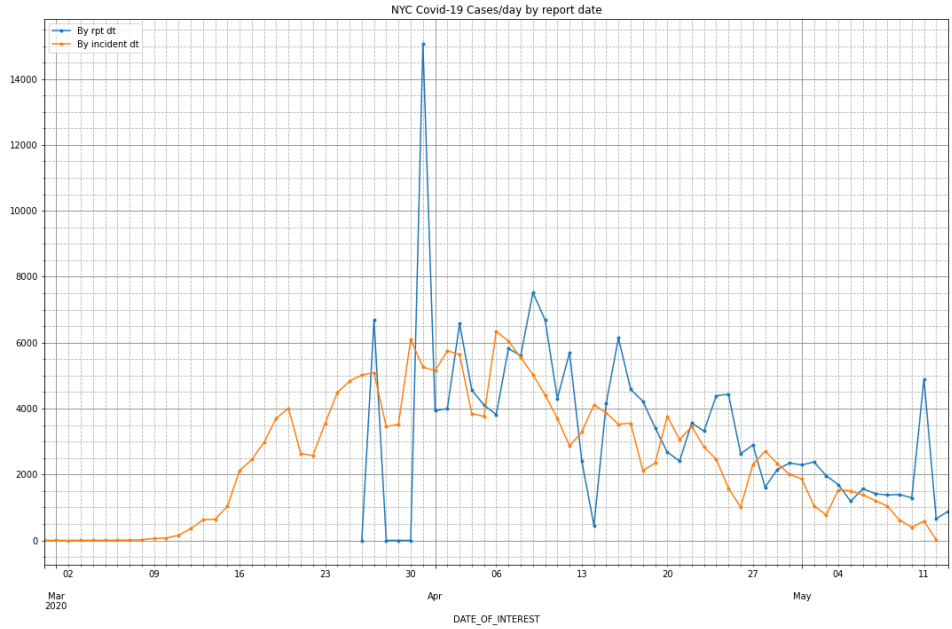

FIGURE 4. NYC cases per day by incident date vs by reporting date.

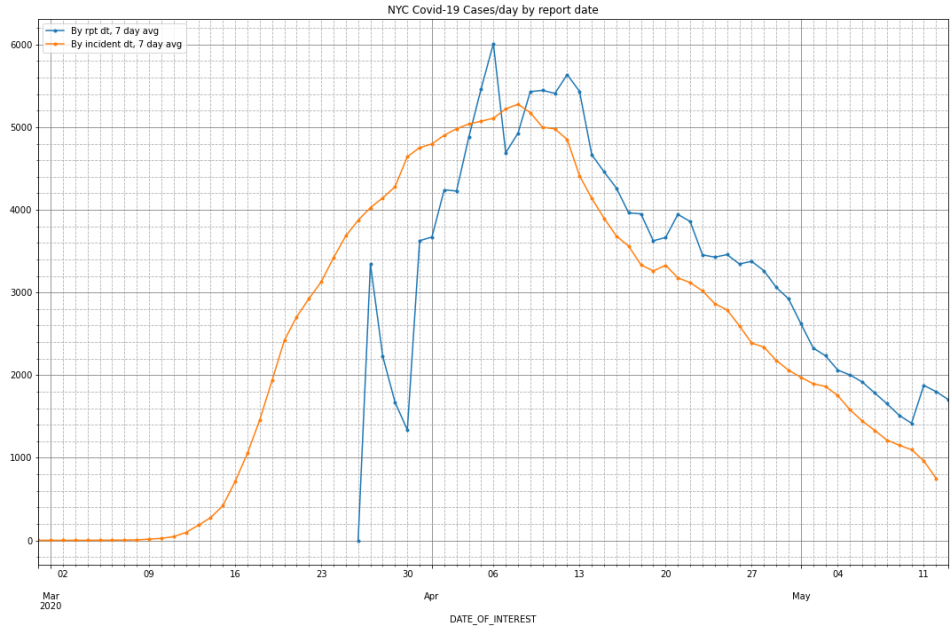

FIGURE 5. NYC cases per day, 7 day rolling average, by incident date vs by reporting date.

As you can see, the reporting date data is far noisier; so much so that the 7 day cycle is obscured and the 7 day rolling window still shows substantial noise. For example, the spike on May 11th in the raw data of about 5,000 cases corresponds to the history being restated slightly, but going back to the third week of March. This gives a substantially distorted view from day to day. The noise also makes the peak appear to have occurred much earlier than it actually does.

The 7 day rolling window shows that the report date data is overstating the 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.

4. Conclusions

During this pandemic, it's great to see that 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's also disturbing that news agencies are reporting on these numbers as if they actually occurred on the reporting date, and governments may be taking action based on the same misconception.

I find it surprising that epidemiologists who make a career out of analyzing epidemics and pandemics would record the data in such a fashion. But, on the other hand, I suppose such work tends to be on a longer time scale and it's only in the current pandemic that we needed accurate, up to date infection and death counts.

I'd hope that someone would take it upon themselves to collect and aggregate the data on an incident date basis. This would be a huge undertaking, but the longer this pandemic persists, the more important this becomes.

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.

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

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

[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.

[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 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.

[Ste20d] 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.

Email address: hjstein@bloomberg.net