COVID-19: 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.

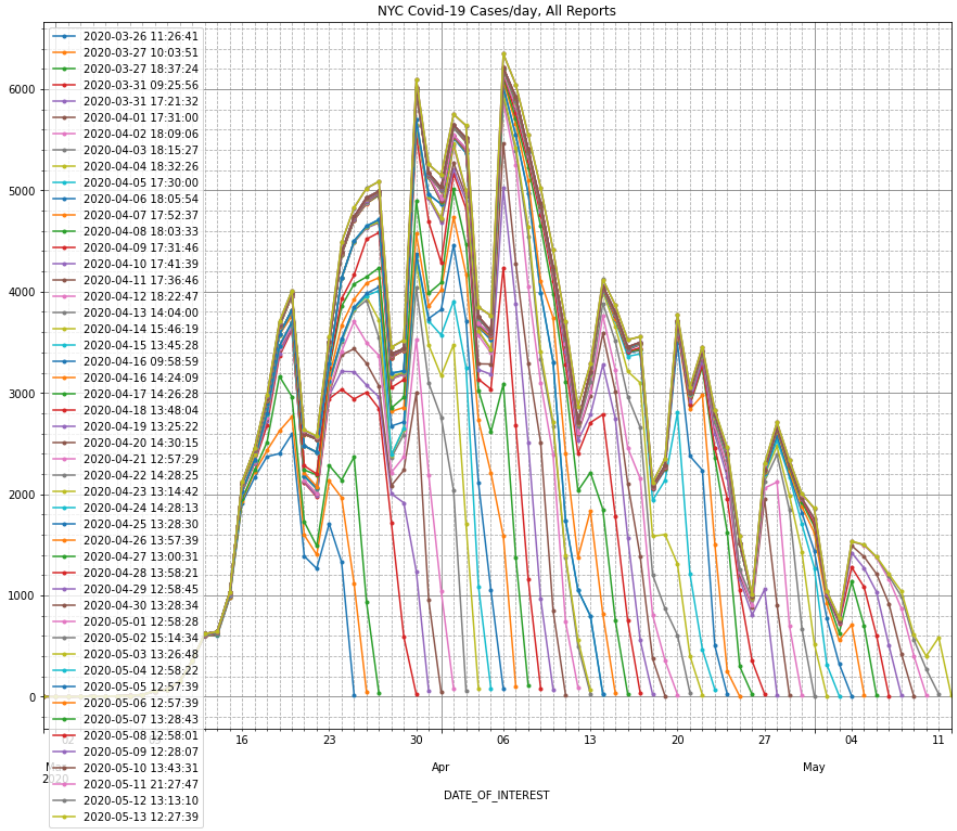

FIGURE 1. NYC COVID-19 cases per day from each daily report. Data from the NYC COVID-19 data github repository.

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.

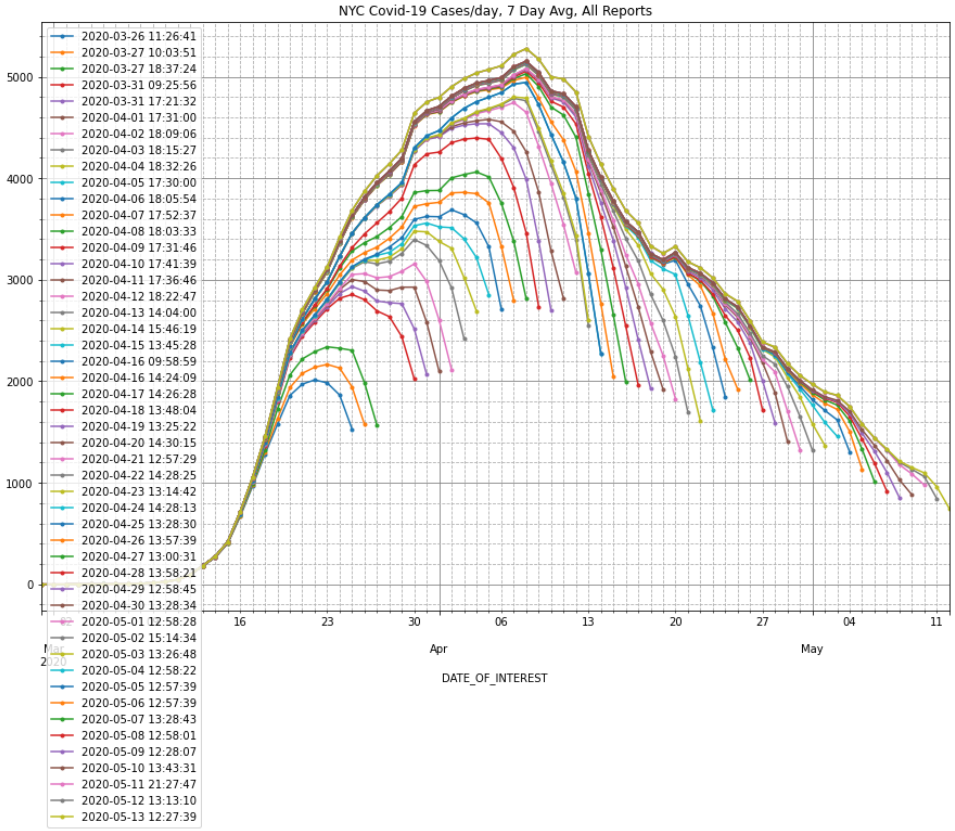

FIGURE 2. 7-day rolling average of NYC COVID-19 cases per day from each daily report. Data from the NYC COVID-19 data github repository.

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

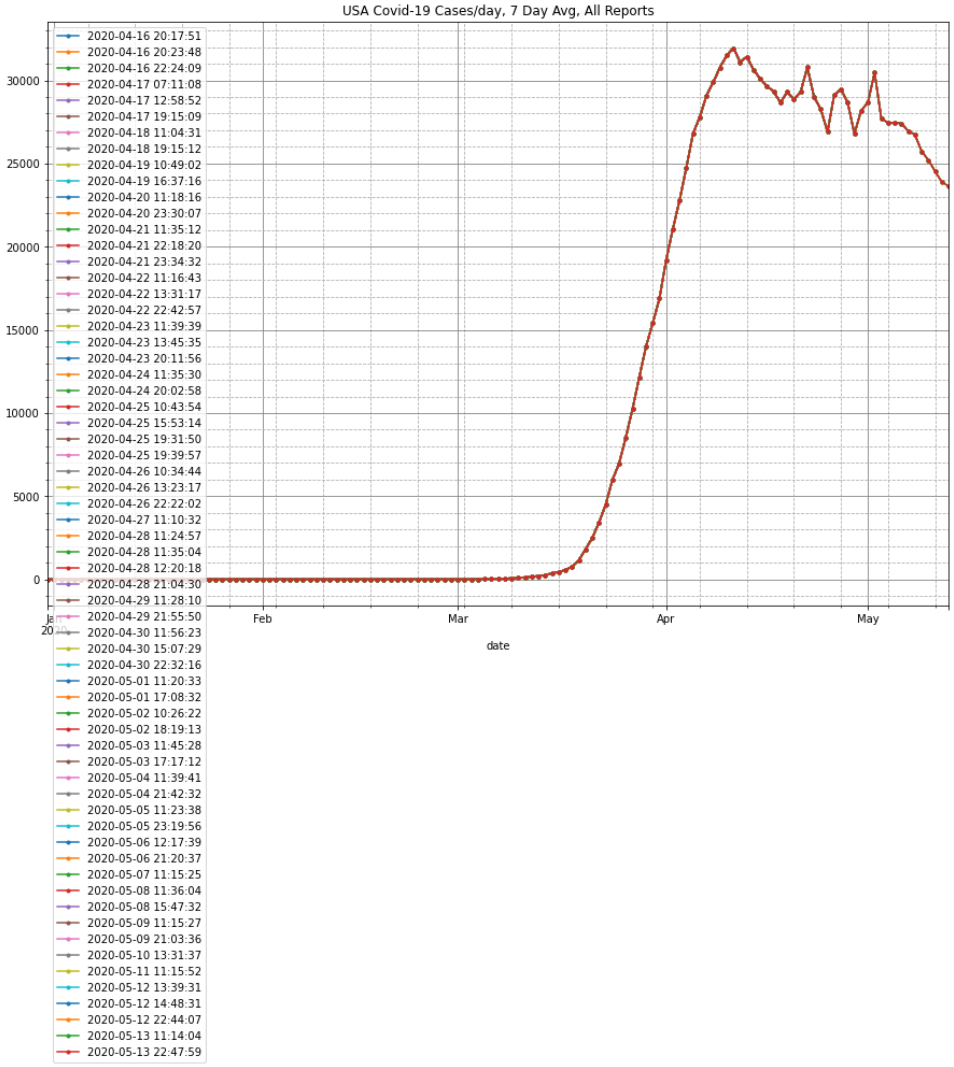

FIGURE 3. USA COVID-19 cases per day from each daily report. Data from the OWID COVID-19 data github repository.

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.

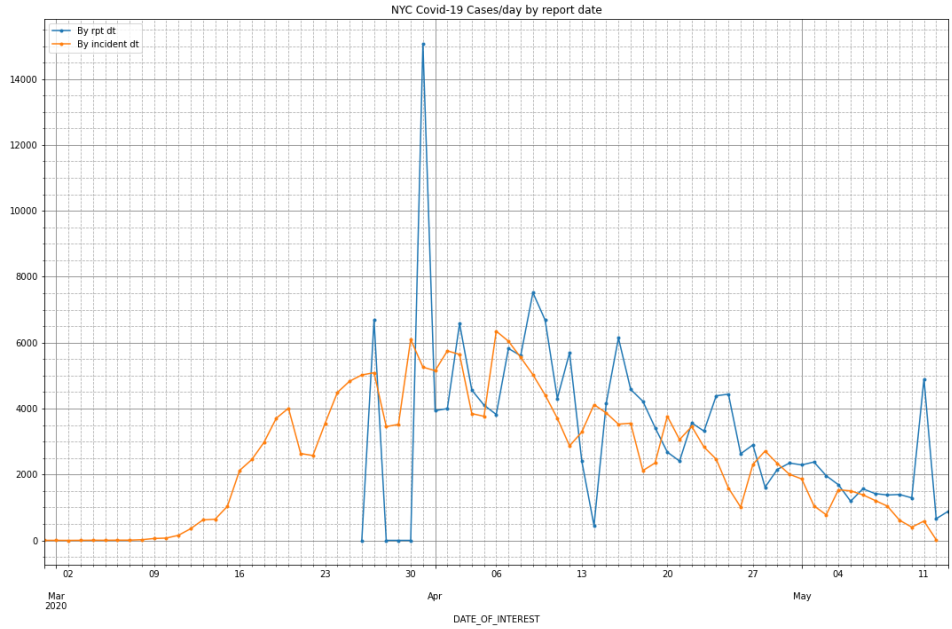

FIGURE 4. NYC cases per day by incident date vs by reporting date.

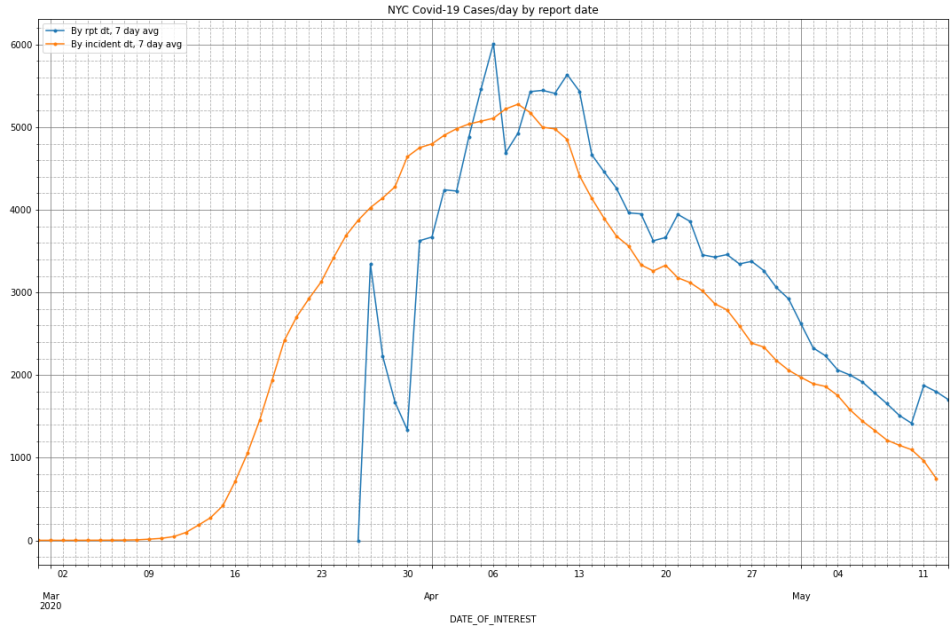

FIGURE 5. NYC cases per day, 7 day rolling average, by incident date vs by reporting date.

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.

Email address: hjstein@bloomberg.net

Bloomberg L.P., New York, NY, USA