COVID-19: GARBAGE IN, GARBAGE OUT

HARVEY J. STEIN

Abstract. COVID-19 data discussion.

Contents

1. Introduction 1

2. Garbage in? 3

3. Garbage out? 5

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.

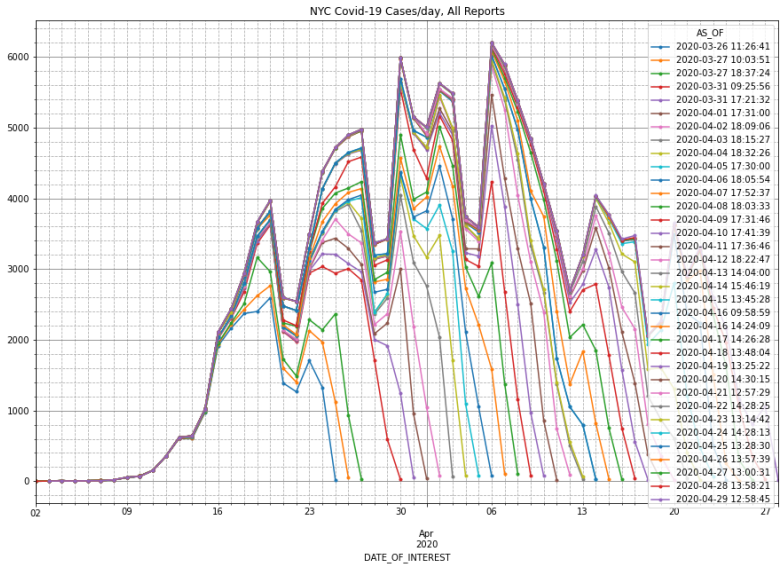

FIGURE 1. NYC COVID-19 cases per day from each daily report. Data from the NYC COVID-19 data github repository.

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.

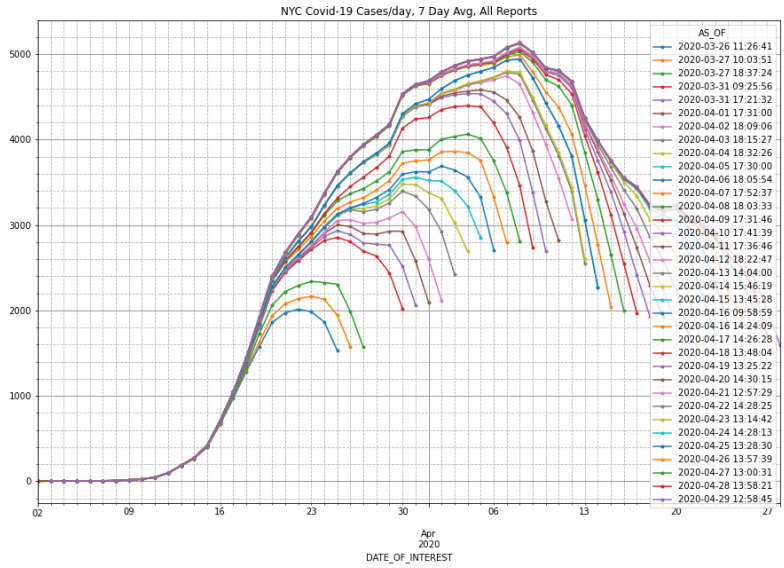

FIGURE 2. 7-day rolling average of NYC COVID-19 cases per day from each daily report. Data from the NYC COVID-19 data github repository.

After seeing this analysis, Jon Asmundsson, editor of *Bloomberg Markets*, wrote back to me:

> Your analysis is really interesting. Have you looked at other states/cities?

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

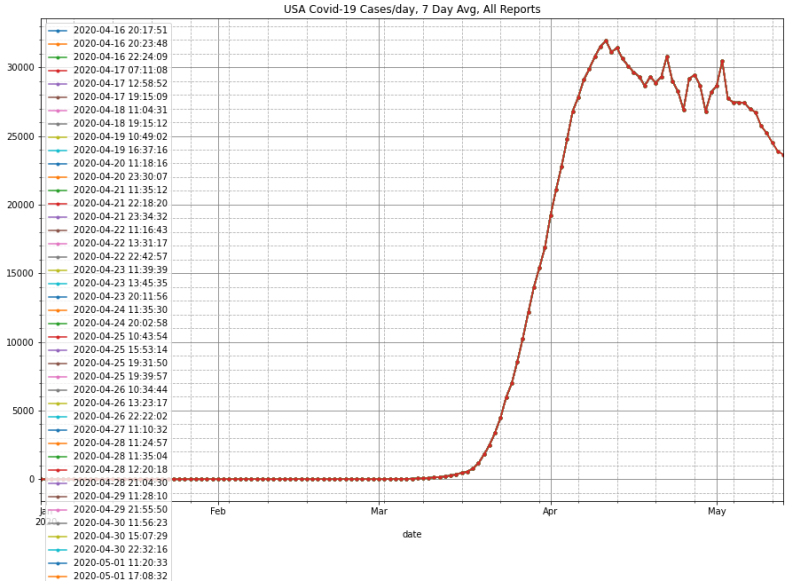

FIGURE 3. USA COVID-19 cases per day from each daily report. Data from the OWID COVID-19 data github repository.

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. He wrote:

> Hi Harvey,
>
> Thanks for getting in touch. There's no perfectly clear and obvious answer but here's how I understand it: it comes down to who is publishing that data and what they consider their "role" to be.
>
> - If they're a government, ministry, public health authority, etc. whose job it is to report on the situation, then they try to give the most accurate picture of how the epidemic evolved in the country, including by going "back in time" and correcting past figures to make them as close as possible to reality.
>
> - On the other hand, organizations like the WHO, ECDC, JHU, consider themselves to be data providers first and foremost. This means they aim for "stability" in their data, and they avoid as much as possible (or even completely) going back and changing past days, as the many people/applications/dashboards/etc. relying on their data wouldn't necessarily notice these changes and handle them correctly.
>
> Another issue is that while it's easy for 1 country to fix past figures in a time series on its own website, it would be much harder for an international organization that receives data from 200+ countries to accept retroactive changes—their job is made a lot easier by simply telling countries to send data corrections as if they were big "blocks" of cases or deaths that suddenly afppear on a given day.
>
> This is obviously an issue sometimes, especially when some of those corrections are very large (for example New York City or China in April), but we don't know of any large and reliable data provider that reports in the way you're looking for.
>
> Best,
>
> Edouard

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 wrote:

> And yes, there are absolutely issues with using reporting data rather than incidence rate, however more often than not, that is all that is available.

3. Garbage out?

So 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. This flattens

the curve. As a result, it makes the infection rate appear lower and it makes the peak appear later. 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.

So how big is the effect, really? I calculated it by taking the NYC data and backing out what it would look like if it was recorded by report date instead of incident date.

Email address: hjstein@bloomberg.net

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