

COVID-19 DATA COLLECTION: GARBAGE IN, GARBAGE OUT

HARVEY J. STEIN

ABSTRACT. It turns out that the major COVID-19 data repositories record cases by report date instead of incident date. We analyze the impact of this.

CONTENTS

1. Introduction	1
2. Garbage in	1
3. Garbage out?	3
4. Conclusions	5
References	6

1. INTRODUCTION

My 4/21 analysis of the NYC COVID-19 data ([COVID-19 NYC Stats – Not What They Seem](#)), shows a graph demonstrating the extent to which missing data has a major impact on the reported incident counts. Figure 1 has the data around the first wave’s peak.

It can take a month or more before all of the data for a given date finally arrives. Moreover, the final count for a given date can be five times larger than the initial count. Figure 2 has the 7 day average report 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 count, 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.

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.

Date: February 1, 2022.

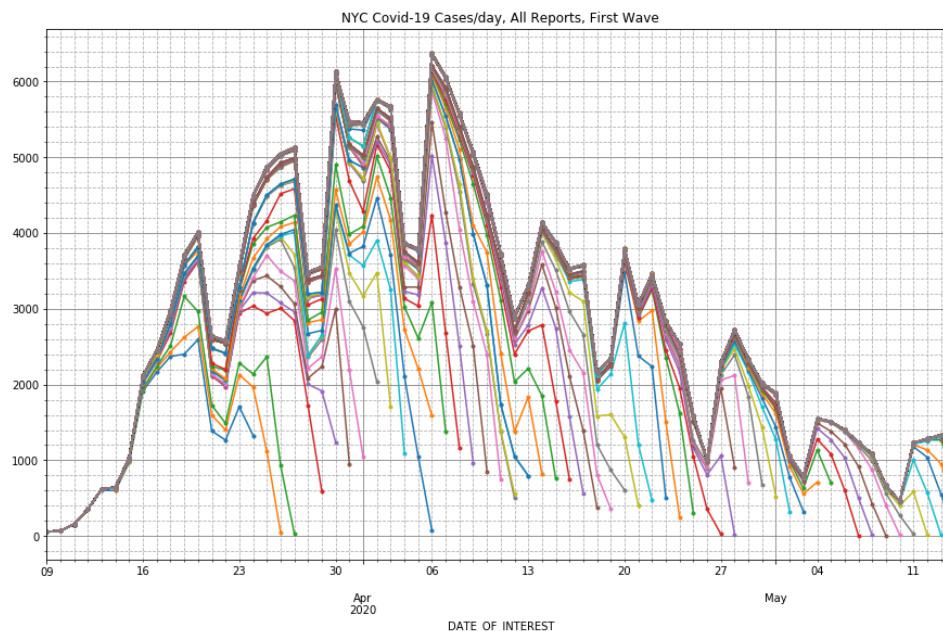


FIGURE 1. NYC COVID-19 cases per day from each daily report around the peak of the first wave. Data from the NYC COVID-19 data github repository.

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 odd must be going on.

So I entered an [issue on the OWID COVID-19 github repository](#). I asked for clarification of 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 large 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.

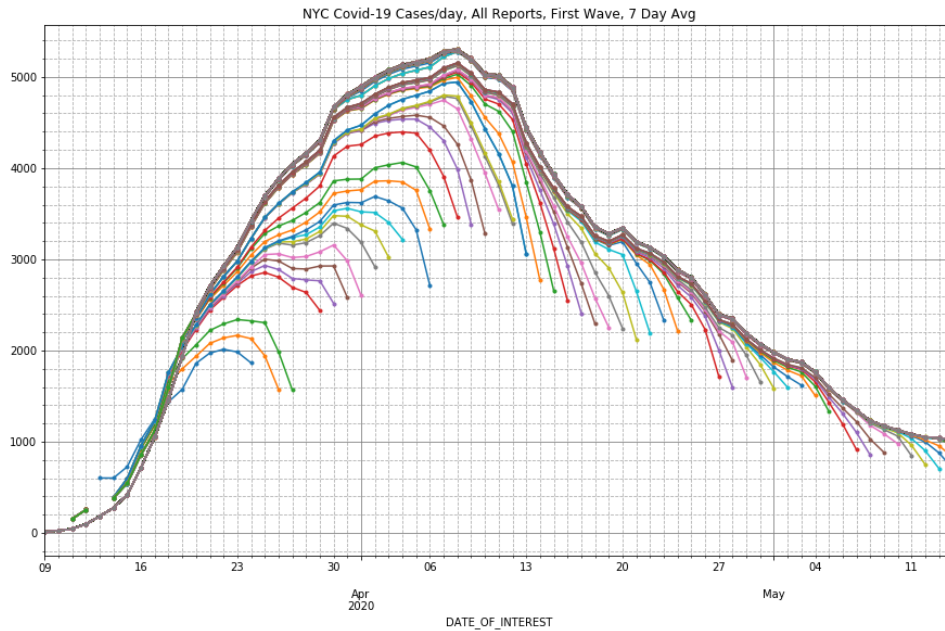


FIGURE 2. 7-day rolling average of NYC COVID-19 cases per day from each daily report. Data from the NYC COVID-19 data github repository.

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 report receipt date instead of by 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 report 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.[\[Gar20\]](#)

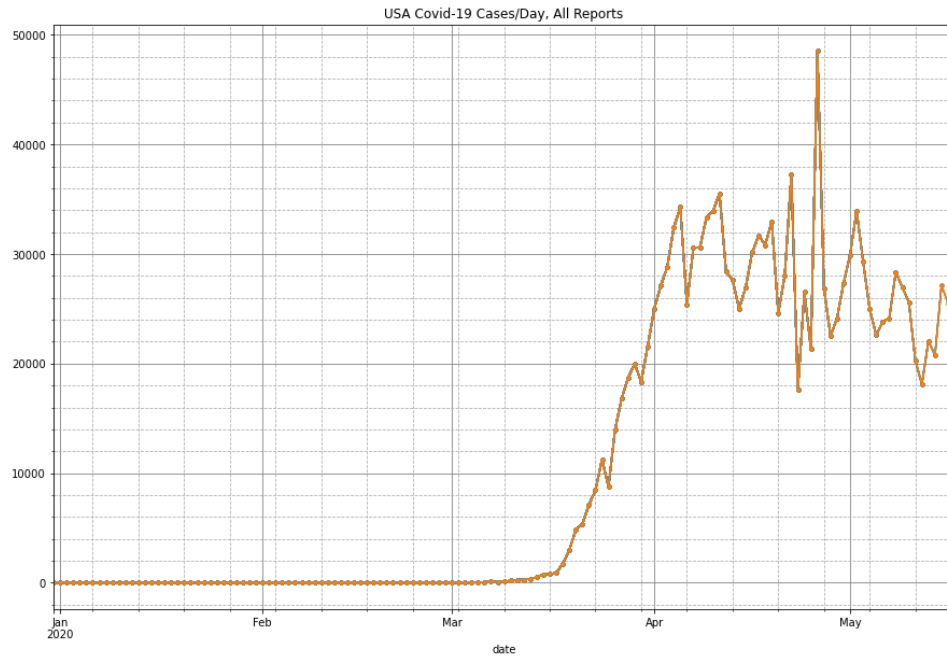


FIGURE 3. USA COVID-19 cases per day from each daily report.
Data from the OWID COVID-19 data github repository.

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 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 ??). It can have a substantial impact.

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 4 and 5 compare the report date results to the incident date results for the first peak and for the Omicron peak.

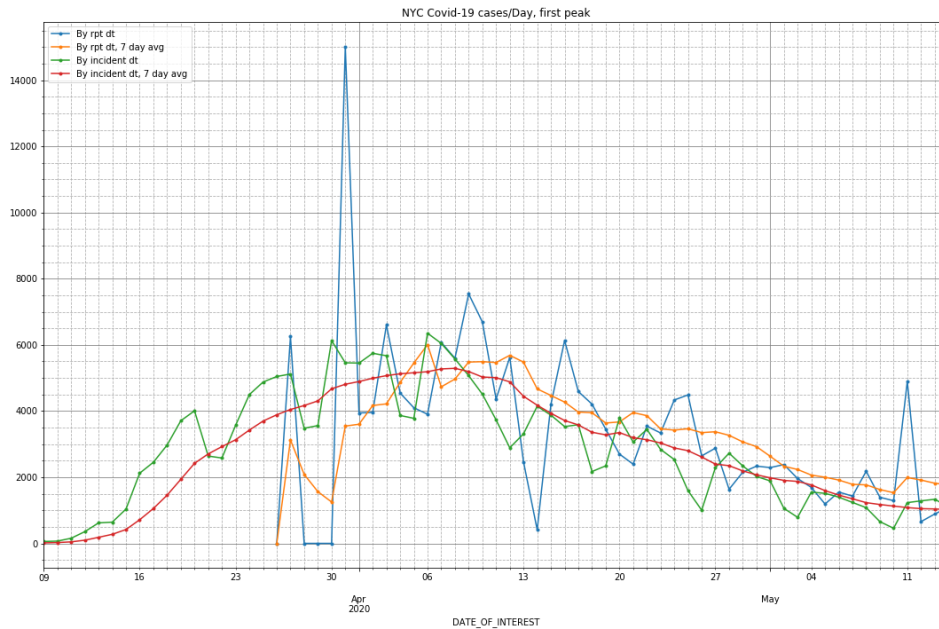


FIGURE 4. NYC cases per day, first peak, by reporting date vs by incident date.

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 The 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.

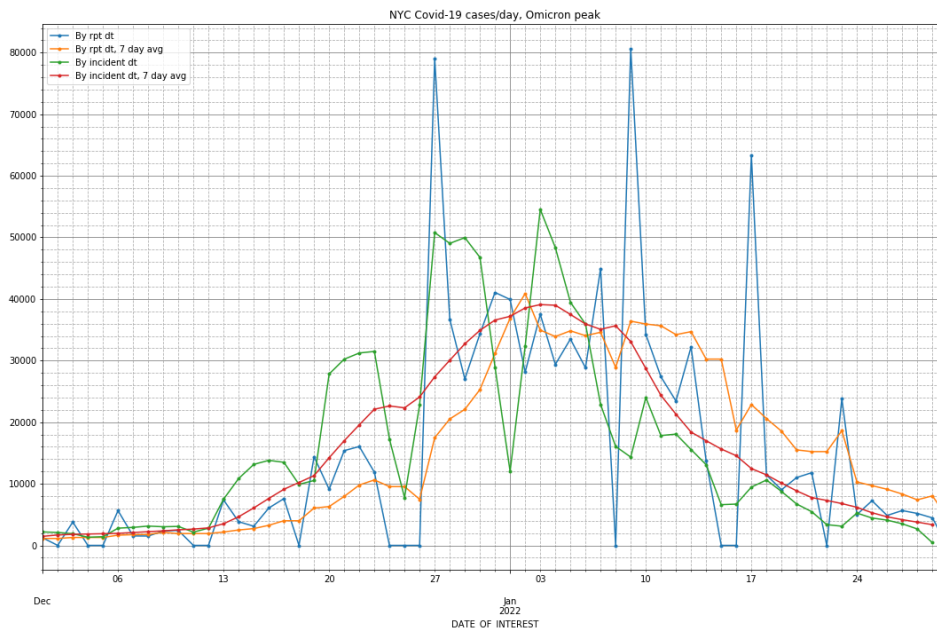


FIGURE 5. NYC cases per day, first peak, by reporting date vs by incident date.

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 hope that data collection agencies will take it upon themselves to collect and aggregate the data on an incident date basis. This would be a significant undertaking, but would better enable us to analyze the data and respond appropriately. 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 further 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 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>.
- [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