COVID-19 Counts in Washington State and Select Other Locations

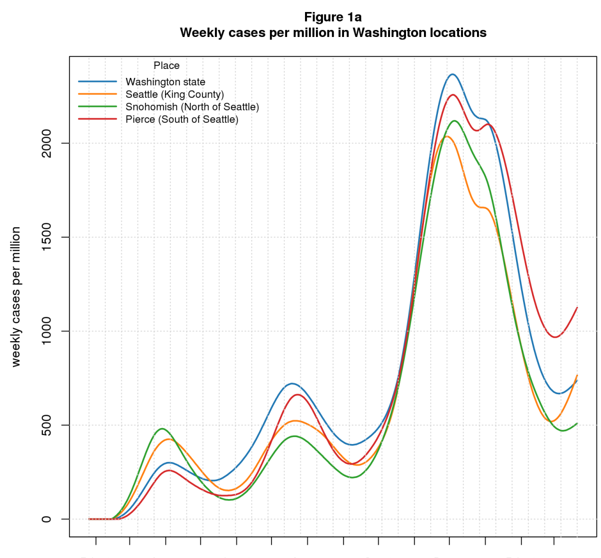
Nathan (Nat) Goodman

March 31, 2021

Counts of cases and deaths are key metrics of COVID-19 prevalence and burden, and are the basis for model-based estimates and predictions of these statistics. I present here graphs showing these metrics over time in Washington state and a few other USA locations of interest to me. I update the graphs and this write-up weekly. Previous versions are here. See below for caveats and details.

Figures 1a-d show *case* counts per million for several Washington and non-Washington locations. The Washington locations are the entire state, the Seattle area where I live, and the adjacent counties to the north and south (Snohomish and Pierce, resp.). The non-Washington locations are Ann Arbor, Boston, San Diego, and Washington DC.

Figures 1a-b (the top row) show smoothed data (see details below); Figures 1c-d (the bottom row) overlay raw data onto the smoothed. The figures use data from Johns Hopkins Center for Systems Science and Engineering (JHU), described below. When comparing the Washington and non-Washington graphs, please note the difference in y-scale: the current Washington rates (about 600-1300 per million in Pierce) are above the rate in San Diego (about 500 per million), overlap the rate in Washington DC (about 1200 per million), and are below the rates in Ann Arbor and Boston (about 2000 per million).



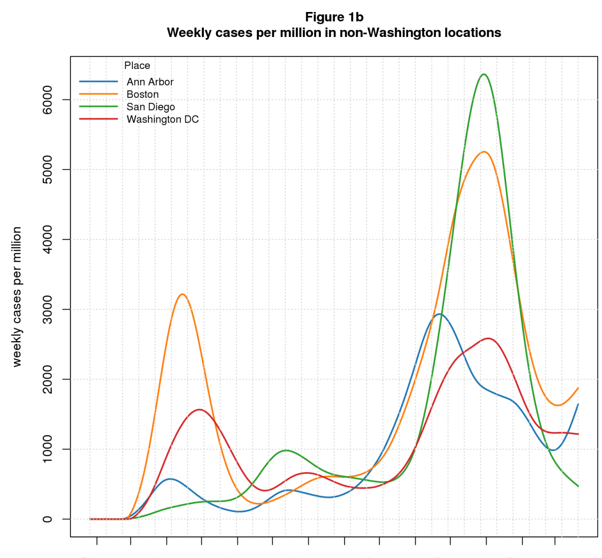


Figure 1c
Weekly cases per million in Washington locations showing raw data

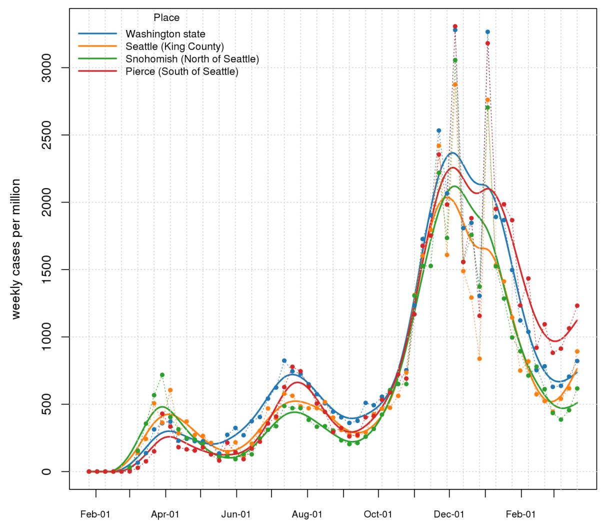
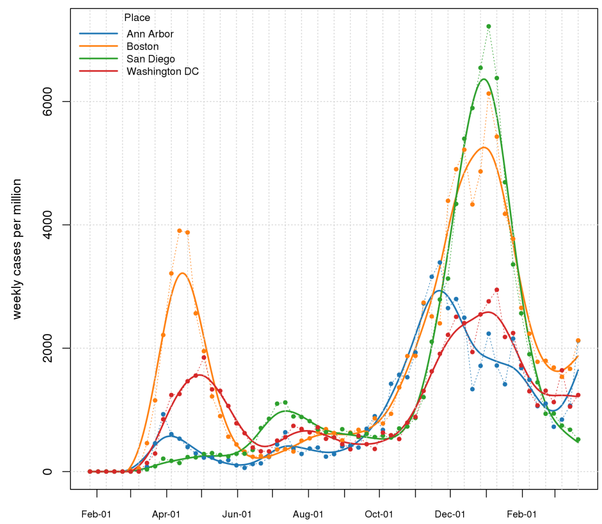


Figure 1d
Weekly cases per million in non-Washington locations showing raw data



The smoothed graphs for Washington (Figure 1a) show that rates, though well below their recent peaks, are starting to climb again in all locations. The raw data (Figure 1c) and simple trend analysis (described below) indicate that the increase is clear in all locations except Snohomish where the increase is not yet definitive. The rate for the state has almost reached its summer peak, while rates for all other locations are well above their summer peaks. Going forward, we can only hope that vaccination of the most vulnerable people will prevent a consequent increase in deaths.

The rates for non-Washington locations (Figures 1b,d) have also fallen dramatically from their peaks but are now heading up in Ann Arbor and Boston, flat in Washington DC, and continuing to decline in San Diego. Trend analysis indicates that the increase in Ann Arbor and decrease in San Diego are real, but the data for Boston and Washington DC are too variable to be confident in the direction.

Figures 2a-d show *deaths* per million for the same locations. When comparing the Washington and non-Washington graphs, again please note the difference in y-scale: the current Washington rates are 4-7 per million; the non-Washington rates are 5-22 per million.

Figure 2a
Weekly deaths per million in Washington locations

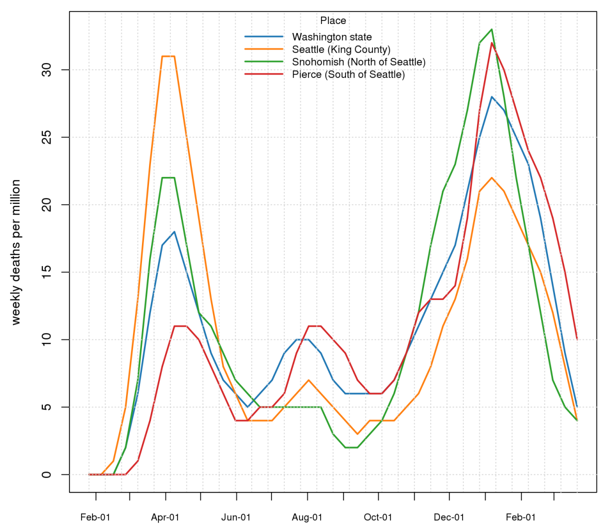


Figure 2b
Weekly deaths per million in non-Washington locations

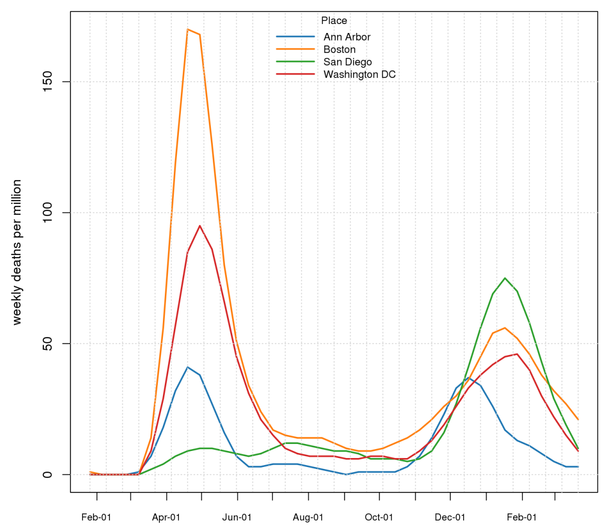


Figure 2c
Weekly deaths per million in Washington locations showing raw data

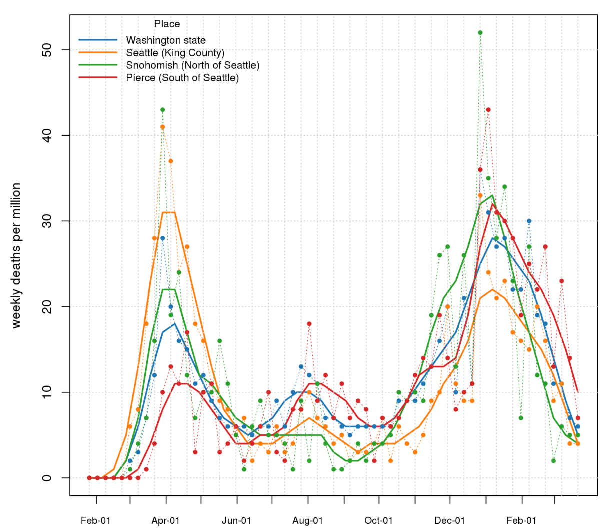
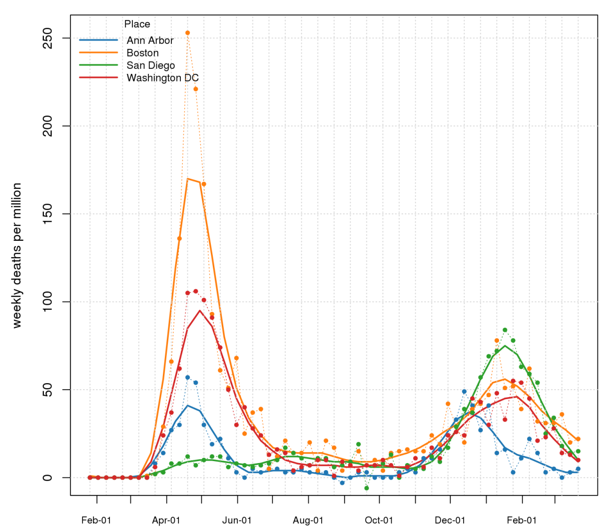


Figure 2d
Weekly deaths per million in non-Washington locations showing raw data



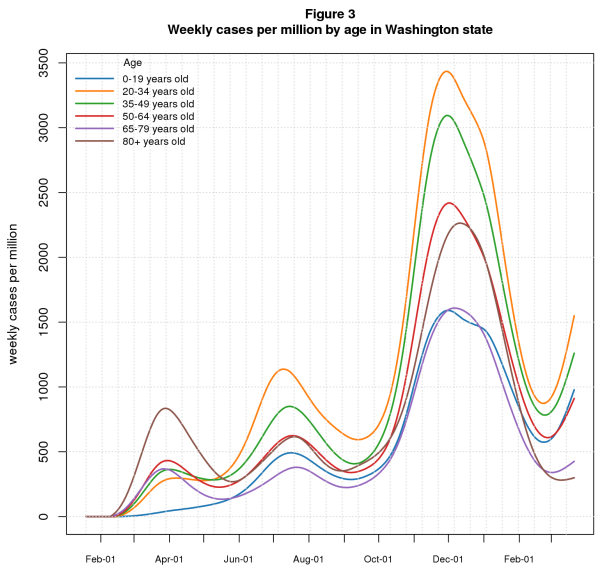
The smoothed Washington data (Figure 2a) shows three waves. The second peak was thankfully lower than the first; the third wave exceeded the first in all areas except Seattle (King County). The graphs are well down from the third peaks and are finally below the second peaks. The raw data (Figure 2c) remains quite variable; trend analysis (described below) indicates that looking back 6 weeks, the decline is clear.

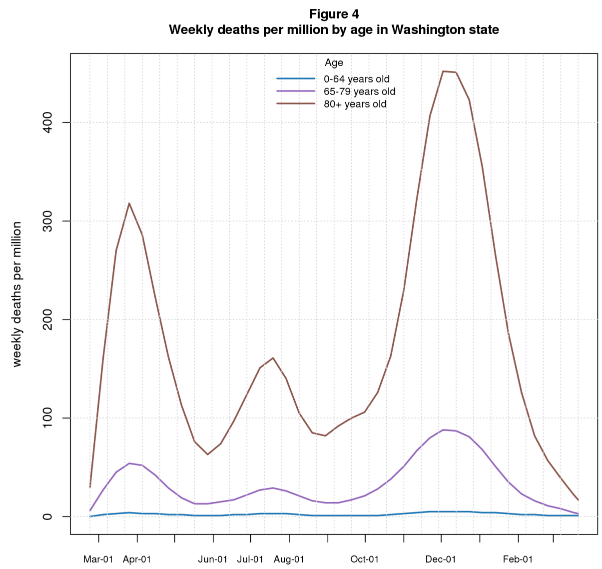
The smoothed non-Washington data (Figure 2b) shows early peaks in most locations, followed by a long trough, followed by a second wave starting in November. The graphs are well down from their recent peaks and seem to be continuing down. The raw data, though variable, broadly supports this view; trend analysis (described below) indicates that looking back 6-8 weeks, the decline is clear.

The next graphs show the Washington results broken down by age. This data is from Washington State Department of Health (DOH) weekly downloads, described below. An important caveat is that the DOH download systematically undercounts events in recent weeks due to manual curation. I extrapolate data for late time points as discussed below.

In most versions before March 17, 2021, I showed these graphs for each Washington location. Here I only show the statewide graphs: the other locations are similar. In past, DOH reported results in 20 year age groups starting with 0-19, with a final group for 80+. As of the March 14, 2021 data release (corresponding to document version March 17), DOH changed age groups. The new groups are 0-19, followed by 15 year ranges (20-34, 35-49, 50-64, 65-79) with a final group for 80+.

Figure 3 is *cases*. Figure 4 is *deaths*; this graph aggregates 0-64 into a single group, since the death rate in these ages is near 0.





Early on, the pandemic struck older age groups most heavily. Over time, *cases* spread into all age groups, even the young. During the second wave, older groups did better in most locations with young adults and middle aged adults (20-34 and 35-49 years) becoming the most affected groups. The third wave swept into all age groups. As the wave grew, the oldest people (80+) were again strongly affected. The curves are now heading back up in all age groups, but thankfully the oldest groups (65-79 and 80+) seem to be less affected.

The shocking devastation of the 80+ age group jumps off the page in Figure 4. The death rate in this age group shows three waves. The third wave seems to have reached its peak but caution is in order because the DOH death seems to be lagging JHU.

Caveats

1. The term *case* means a person with a detected COVID infection. In some data sources, this includes "confirmed cases", meaning people with positive molecular COVID tests, as well as "probable cases". I believe JHU only includes "confirmed cases" based on the name of the file I download.
2. Detected *cases* undercount actual cases by an unknown amount. When testing volume is higher, it's reasonable to expect the detected count to get closer to the actual count. Modelers attempt to correct for this. I don't include any such corrections here.
3. The same issues apply to *deaths* to a lesser extent, except perhaps early in the pandemic.
4. The geographic granularity in the underlying data is *state* or *county*. I refer to locations by city names reasoning that readers are more likely to know "Seattle" or "Ann Arbor" than "King" or "Washtenaw".
5. The date granularity in the graphs is weekly. The underlying JHU data is daily; I sum the data by week before graphing.

6. I truncate the data to the last full week prior to the week reported here.

Technical Details

1. I smooth the graphs using a smoothing spline (R's `smooth.spline`) for visual appeal. This is especially important for the *deaths* graphs where the counts are so low that unsmoothed week-to-week variation makes the graphs hard to read.
2. The Washington DOH data (used in Figures 3 and 4 to show counts broken down by age) systematically undercounts events in recent weeks due to manual curation. I attempt to correct this undercount through a linear extrapolation function (using R's `lm`). I have tweaked the extrapolation repeatedly, even turning it off for a few weeks. The current version uses a model that combines date and recentness effects. In past, I created models for each Washington location and age group but had to change when DOH changed its age groups on March 14, 2021 (see below). I now create a single model for the state as a whole and all age groups summed together, then blithely apply that model to all locations and ages.
3. The trend analysis computes a linear regression (using R's `lm`) over the most recent four, six, or eight weeks of data and reports the computed slope and the p-value for the slope. In essence, this compares the trend to the null hypothesis that the true counts are constant and the observed points are randomly selected from a normal distribution. After looking at trend results across the entire time series, I determined that p-values below 0.1 indicate convincing trends; this cutoff is arbitrary, of course.

Data Sources

Washington State Department of Health (DOH)

DOH provides three COVID data streams.

1. Washington Disease Reporting System (WDRS) provides daily "hot off the presses" results for use by public health officials, health care providers, and qualified researchers. It is not available to the general public, including yours truly.
2. COVID-19 Data Dashboard provides a web graphical user interface to summary data from WDRS for the general public. (At least, I think the data is from WDRS - they don't actually say).
3. Weekly data downloads (available from the Data Dashboard web page) of data curated by DOH staff. The curation corrects errors in the daily feed, such as, duplicate reports, multiple test results for the same incident (e.g., initial and confirmation tests for the same individual), incorrect reporting dates, incorrect county assignments (e.g., when an individual crosses county lines to get tested).

The weekly DOH download reports data by age group. In past, the groups were 20-year ranges starting with 0-19, with a final group for 80+. As of March 14, 2021, they changed the groups to an initial 20-year range (0-19), then several 15 year ranges (20-34, 35-49, 50-64, 65-79), with a final group for 80+. My code cannot yet handle this change.

Johns Hopkins Center for Systems Science and Engineering (JHU)

JHU CSSE has created an impressive portal for COVID data and analysis. They provide their data to the public through a GitHub repository. The data I use is from the csse_covid_19_data/csse_covid_19_time_series directory: time_series_covid19_confirmed_US.csv for *cases* and time_series_covid19_deaths_US.csv for *deaths*.

JHU updates the data daily. I download the data the same day as the DOH data (now Tuesdays) for operational convenience.

Other Data Sources

I use two other COVID data sources in my project although not in this document.

- New York Times COVID Repository. The file I download is us-counties.csv. Like Washington DOH and JHU, NYT has county-level data. Unlike these, it includes "probable" as well as "confirmed" cases and deaths; I see no way to separate the two categories.
- COVID Tracking Project. This project reports a wide range of interesting statistics (negative test counts, for example), but I only use the *case* and *death* data. It does not provide county-level data so is not useful for the non-Washington locations I show. The file I download is https://covidtracking.com/data/download/washington-history.csv. I use this only as a check on the state-level Washington data from the other sources. As of March 7, 2021, the project ended its data collection efforts.

The population data used for the per capita calculations is from Census Reporter. The file connecting Census Reporter *geoids* to counties is the Census Bureau Gazetteer.

Comments Please!

Please post comments on Twitter or Facebook, or contact me by email natg@shore.net.