Analysis of Covid 19 data for US and North Carolina

I have analyzed available covid 19 data through 3 May 2020. Sources for this analysis are: (1) data on number of new cases, cumulative number of cases, and new deaths, for US and NC- https://coronavirus.1point3acres.com/en; (2) data on daily number of tests in US- https://ourworldindata.org/grapher/full-list-total-tests-for-covid-19?time=2020-03-14.. ; (3) data on daily number of new cases per test for NC- https://www.ncdhhs.gov/divisions/public-health/covid19/covid-19-nc-case-count.

I make some projections here about when the first wave of infections will have passed. These projections are not based on epidemiological models. Rather, they are based on statistical models. In particular, I fit non-linear regressions to the current data and make projections based on the best fit regression equation. It should be kept in mind that such projections assume that the processes that have been occurring up to now will continue on into the future. That assumption, of course, might not be true.

I. Number of new cases in US.

The number of new cases per day in the US has leveled off and been at a plateau since about 1 Apr. (Fig. 1). Because the number of tests per day has increased steadily increased in this period (Fig. 2), one would expect to see an increase in the number of new cases per day if the true number of new cases were steady. Since we don’t see that, it suggests that one needs to correct the number of new cases by dividing by the number of tests.

This can be explained by a simple example. Suppose on day 1 there were 1,000 tests and 100 of them came back positive, while on day 2 there were 2,000 tests and 100 of them came back positive. Assuming that there could have been 2,000 tests done on day 1, we would expect to see 200 positive cases on that day (twice as many as observed because twice as many tests). In that case, instead of seeing the same number of new cases on the two days, we would have seen 200 on day 1 and 100 on day 2.

Another way of looking at this is to calculate the proportion of tests that are positive (which I will call pp for “proportion positive”), which is just the number of new cases/number of tests. For day 1, that would be 100/1,000 = 0.1, whereas for day 2 it would be 100/2,000 = 0.05. In other words, this measure says that there were twice as many new cases on day 1 than on day 2, which is the same result as above. One can thus use the value pp to follow the relative number of new cases over time. It’s ok to use relative numbers, because what we’re interested in here is the time course of infections.

When this is done, it is evident that the number of new cases per test has been declining since the beginning of April (Fig. 3; y-axis is equivalent to pp). Moreover the cumulative pp fits a modified logistic growth curve exceedingly well. This curve is

$$pp = a + K/(1 + e^{-r(day-b)})$$
Fitting this curve using non-linear regression produces Fig. 4, in which the circles are the data and the line is the fitted curve. Extrapolating this curve to 100 days gives a prediction about when the number of new cases should decline to 0 (Fig. 5). The red line indicates the pp value at which the cumulative number of cases per test reaches 95% of the projected total. This curve intersects the fitted curve at day 78, which corresponds to 17 May. This means that 95% of all infections from the first wave of infections is projected to have occurred by this data. Of course, this still means that there will be a large number of new infections after that data. Given a minimum estimate from various sources that more than 1,000,000 individuals in the US will be infected during the first wave, this number is 0.05 x 1,000,000 or 50,000 new infections after 17 May. This is, of course a lot, but a “small” number compared to the total number infected. The corresponding date on which 99.9% of all infections have occurred is 18 June. It is projected that there will be only about 750 new infections after that date. Of course, this assumes that relaxing social distancing restrictions will not trigger a new wave of infections.

II. Number of new deaths in US.

The number of new deaths per day (Fig. 6), which is not affected by the intensity of testing, has plateaued for the last 3 weeks and does not yet show any signs of declining.
North Carolina is showing some similar trends for infections (death data are too sparse to analyze statistically). The number of new cases recorded per day continues to rise (Fig. 7), but this seems likely to be due to an increase in the number of tests per day, since the number of new cases per test (pp) continues to decline (Fig. 8). The cumulative new cases per test fits a modified logistic growth curve very well (Fig. 9), and the projection from that curve indicates NC should reach 95% of cases in the first wave on 14 May. 99.9% of all infections are projected to have occurred by 18 June.
Fig. 9

Fig. 10

One caveat is that this analysis assumes the probability of detecting an infection with a test remains constant as the number of tests increases. One possibility to keep in mind is that early on, tests were given primarily to those who were very sick or who were suspected to have come into contact with someone carrying the virus. It is possible that as tests become more available and anyone can get a test at any time, the probability that a test will yield a positive result could decline because a greater proportion of people being tested are less likely to have come into contact with the virus. My suspicion, though, is that the amount of testing that has been done so far has not reached the point at which this effect may become more than negligible.