Title: Accounting for incomplete testing in the estimation of epidemic parameters
Authors: Rebecca A. Betensky and Yang Feng

Abstract:
As the COVID-19 pandemic spreads across the world and the United States, it is important to understand its evolution in real time and at regional levels. The field of infectious diseases epidemiology has highly advanced modeling and estimation strategies that yield relevant estimates. These include the doubling time of the epidemic, i.e., the number of days until the number of cases doubles, and various representations of the number of cases over time, including the epidemic curve and associated cumulative incidence curve. While these quantities are immediately estimable given current data, they suffer from dependence on the underlying testing strategies within communities. Specifically, they are inextricably tied to the likelihood that an infected individual is tested and identified as a case. We clarify the functional relationship between testing and the epidemic parameters of interest, and thereby derive sensitivity analyses that explore the range of possible truths under various testing dynamics. We demonstrate that crude estimates that assume stable testing or complete testing can be overly-optimistic.

Introduction:
The features of an epidemic are summarized and visualized using several measures. One such measure is the epidemic curve[1], which depicts numbers of reported cases as a function of time. Another is the cumulative version of this curve, which depicts total reported cases as a function of time. While these are important and of high interest for the purpose of understanding the dynamics of the epidemic and the utility of public health interventions, they are limited by the testing processes. That is, the numbers of cases identified are limited by the numbers of tests conducted. A decrease or flattening of the epidemic curve could be due to a decrease in the rate of infection, or it might be due to a decrease in testing, or some combination of the two. An alternative measure of the epidemic is its current *doubling time*[2], i.e., the time needed for the cumulative incidence to double. We suggest that the doubling time should be used as a primary measure of the epidemic due to its clear interpretation in light of testing policies and dynamics and the potential to conduct meaningful sensitivity analyses of it. Likewise, cumulative incidence curves that are normalized to recent dates share the same desirable features.

If the epidemic follows an exponential growth model, the doubling time is constant. That is, the time for the number of cases to double remains the same at all times during the course of the epidemic. An increase in the doubling time is an indicator that the growth of the epidemic is slowing, which in turn, indicates that public health policies, such as social distancing, are displaying efficacy. For this reason, the doubling time is a useful descriptor of the epidemic.

It is important to understand how changes in testing policies and implementation might affect the estimates of doubling time, given that as for the epidemic curve, it is intertwined with testing. The true doubling time at time *t* is defined as

$$T_d = \max\left\{d: \frac{C(t)}{C(t-d)} \leq 2\right\},$$

where *C(t)* is the actual total number of cases at time *t*. Because we cannot know C(t), instead we estimate the doubling time at time t as

$$\tilde{T}_d = \max\left\{d: \frac{O(t)}{O(t-d)} \leq 2\right\},$$

where O(t) is the observed total number of cases at time t. It is necessarily the case that O(t) ≤C(t). There is a simple relationship between O(t) and C(t), which follows from an application of Bayes theorem:

$$Prob(infected \ \leq \ t) = \frac{Prob(infected \ \leq \ t|tested \leq t) \times Prob(tested \leq t)}{Prob(tested \leq t|infected \leq t)}$$

Or equivalently,

$$Prob(infected \leq t) = \frac{Prob(infected \ \leq t \ and \ tested \leq t)}{Prob(tested \leq t|infected \leq t)}.$$

This implies that we can estimate the true number infected, C(t), using its expected value, E(C(t)), which is given by:

$$E\big(C(t)\big) = \frac{O(t)}{Prob(tested \leq t|infected \leq t)}.$$

This, in turn, means that we can re-express the true doubling time as a function of the observed numbers of cases, along with the probabilities of testing of those who are infected:

$$T_d \cong \max\left\{d\colon \frac{E(C(t))}{E(C(t-d))} \ \leq 2\right\} = \max\left\{d\colon \frac{O(t)}{O(t-d)} \times R(t,d) \ \leq 2\right\},$$

where

$$R(t,d) = \frac{Prob(tested \leq t-d|infected \leq t-d)}{Prob(tested \leq t|infected \leq t)}.$$

This expression clarifies the limitations in estimating the true doubling time; since we do not know the proportions of infected individuals who are tested at different times, we do not know R(t,d). Nonetheless, it provides us with an understanding of what precisely the estimated doubling time is, and that it is a good estimate of the true doubling time when (t,d) are such that R(t,d) is approximately equal to one. For example, it is reasonable to expect that for t large enough (i.e., enough time into the epidemic) and for d small enough (i.e., short intervals of time), the probability that an infected individual is tested is constant. In particular, if the probability of testing of infected is constant over the d units of time in (t-d,t), this renders the observed doubling time estimate an accurate estimate of the true doubling time.

In addition to assuming the constancy of R(t,d), we can conduct sensitivity analyses to determine the robustness of this assumption. In particular, we could assume that in the recent past, the probability of testing of infected might have decreased on day t-s, but otherwise remained constant. This is a conservative assumption since its effect is to increase R(t,u) above 1 for u≤(s-1) and for it to remain equal to 1 for u>s.

An alternative, exploratory analysis is a visualization of the cumulative incidence curve, anchored to a recent date, t*, such as 9 days prior to the current date. Using the observed counts for this amounts to a plot of O(t)/O(t*) versus t, for t>t*. Using the same reasoning as above,

$$\frac{C(t)}{C(t^*)} \cong \frac{E\big(C(t)\big)}{E\big(C(t^*)\big)} \cong \frac{O(t)}{O(t^*)} \times R(t, t-t^*).$$

If $R(t, t-t^*) \cong 1$, the observed relative cumulative incidence curve provides a good estimate of the true relative epidemic curve. If this assumption is not plausible within (t*,t), then alternative values for $R(t, t-t^*)$ can be used in sensitivity analysis.

We have estimated the doubling times and cumulative incidence curves nonparametrically. Parametric alternatives are possible, ranging from log-linear Poisson regression with offset terms to account for testing probability estimates, to fully specified epidemic models. The nonparametric approach is simplest and appropriate when it is desirable that the data fully guide the estimation. An imposed linear assumption has immediate consequences on estimation, and may or may not be accurate.

Results:

Using current data on the number of positive tests in the United States and territories (covidtracking.com) we have estimated the current (April 7, 2020) doubling times for the 12 states with the most cases (Figure 1). The open circle indicates the estimate under the assumption that the probability of testing infecteds has not changed over the past few days (2-8 for the states depicted). We linearly interpolated the discrete-time estimates. The curves depict the current doubling time under the scenario that the probability of testing of infecteds was constant in the past, subsequently decreased on a single day in the past, and sustained that decrease through the current time. The decreases depicted are 10% (black curve), 20% (red curve) and 25% (green curve). Of note, the current estimates of doubling time that do not account for potential changes in testing of infecteds might overestimate the true doubling time by as many as three days.

In Figure 2, we display the interpolated doubling times as a function of day to show how they have increased over time. Again, these estimates display sensitivity to potential changes in testing of infected individuals. The black curves assume stable probability of testing over the 20 days considered, the red curves assume a decrease of 10% in probability of testing infected individuals on each current day versus past days, the green curves assume a decrease of 20% and the blue curves assume a decrease of 25%. This provides another view, over time, of the potential for overly optimistic estimates of doubling time if testing is not considered.

In Figure 3, we display the cumulative incidence curves, standardized by the number of cases identified 9 days ago on March 29, 2020. The black curve assumes that testing of infected individuals has not changed in this timeframe. The colored curves represent increases in the probability of testing of infecteds on each current day. These curves demonstrate that the standardized epidemic curve is subject to underestimation, as a function of testing probabilities.

Discussion:

In summary, we have illustrated precisely how the nature of testing among infected individuals affects the estimation of important epidemic parameters, which are used to evaluate the utility of public health interventions in communities. In fact, in New York City, the testing of infecteds does not appear to have changed much at all over the past month (D. Kudlowitz, personal communication). This may not be the case in other states or regions. For this reason, it is important to consider such sensitivity analyses to temper enthusiasm about decreasing doubling

times and flattened epidemic curves to appropriately evaluate the effects of public health interventions.

References:

Data were downloaded from covidtracking.com on April 7, 2020.

1 Wilson, E. B., & Burke, M. H. (1942). The epidemic curve. Proceedings of the National Academy of Sciences of the United States of America, 28(9), 361.

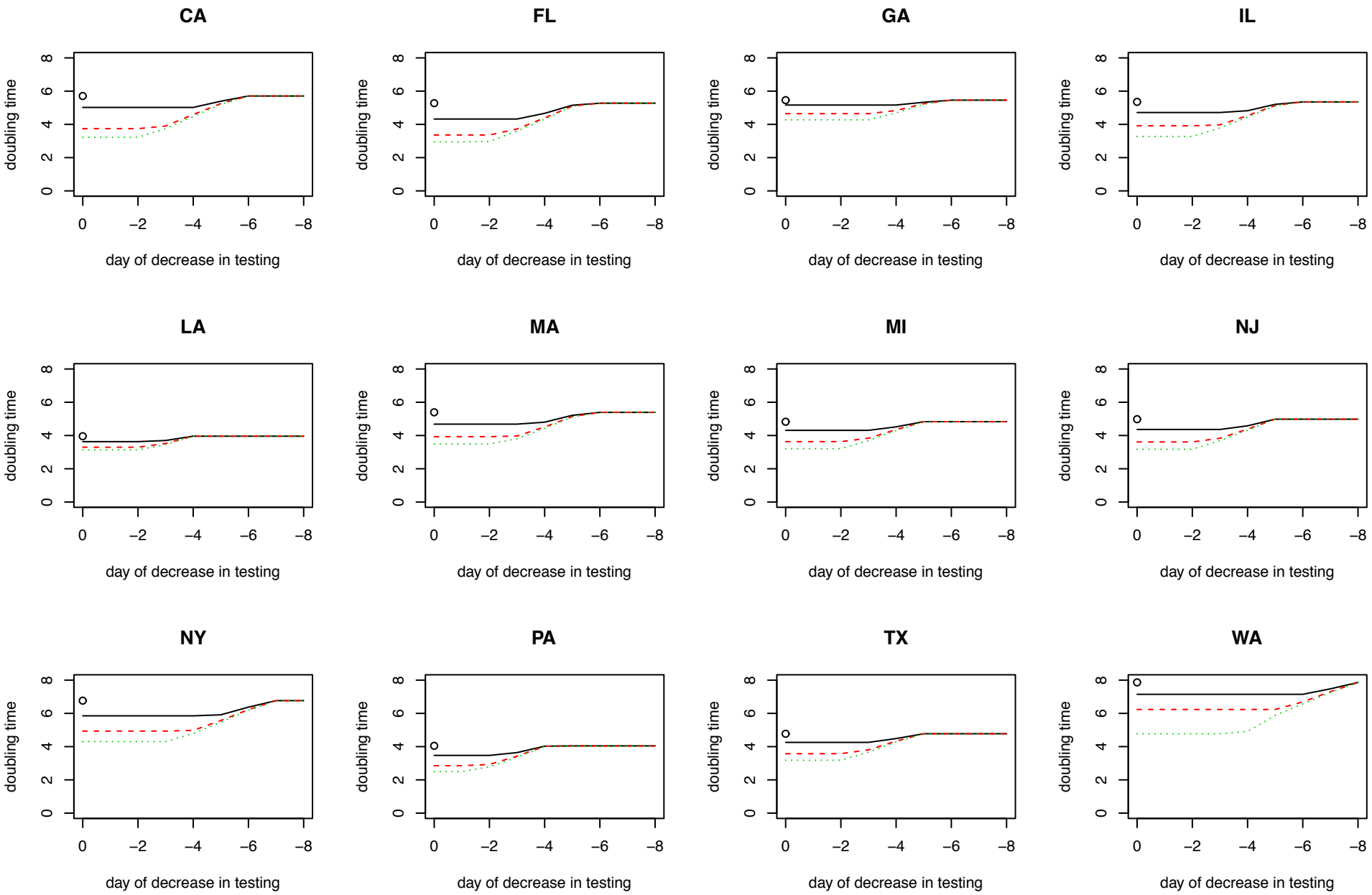
2 Nunes-Vaz, R., 2020. Visualising the doubling time of COVID-19 allows comparison of the success of containment measures. *Global Biosecurity*, 1(3).

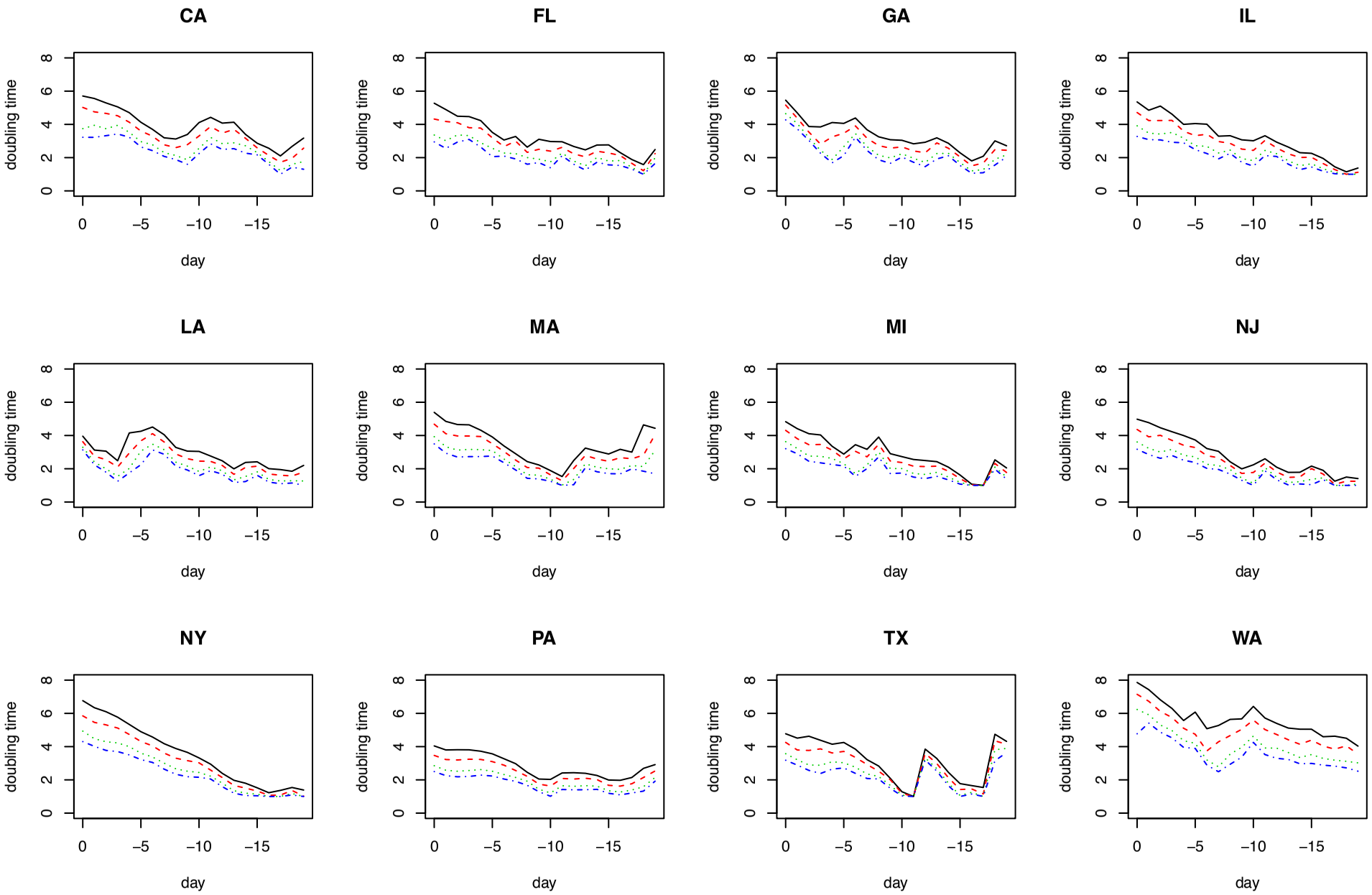
Figure Captions:

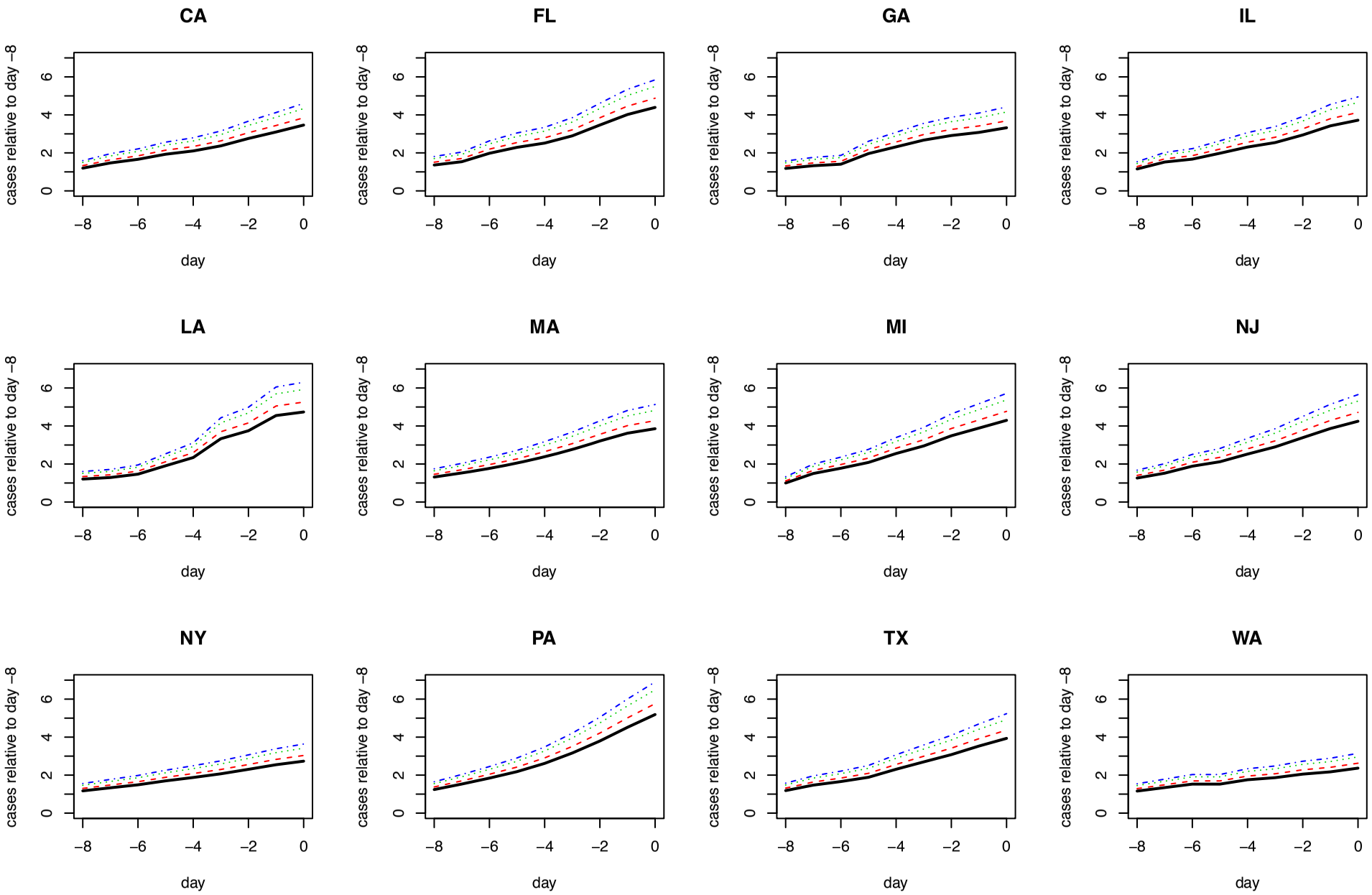
Figure 1: For the 12 states with the most cases as of April 7, 2020 ("day 0"): the open circle is the doubling time estimate on April 7, 2020. The curves depict the current doubling time under the scenario that the probability of testing of infecteds was constant in the past, subsequently decreased on a single day in the past, and sustained that decrease through the current time. The x-axis represents the day on which the probability of testing decreased. The decreases depicted are 10% (black curve), 20% (red curve) and 25% (green curve).

Figure 2: Doubling time on each day from April 7, 2020 back to March 17, 2020. The black curves assume stable probability of testing over the 20 days considered, the red curves assume a decrease of 10% in probability of testing infected individuals on each current day versus past days, the green curves assume a decrease of 20% and the blue curves assume a decrease of 25%.

Figure 3: Cumulative incidence curves, standardized by the number of cases identified 9 days ago on March 29, 2020. The black curve assumes that testing of infected individuals has not changed in this timeframe. The red curves assume a decrease of 10% in probability of testing infected individuals on each current day versus March 29, 2020, the green curves assume a decrease of 20% and the blue curves assume a decrease of 25%.


CA
FL
GA
IL
LA
MA
MI
NJ
NY
PA
TX
WA
doubling time
day of decrease in testing
0
2
4
6
8
0
−2
−4
−6
−8


CA
FL
GA
IL
LA
MA
MI
NJ
NY
PA
TX
WA
doubling time
day
0
2
4
6
8
0
−5
−10
−15


CA
FL
GA
IL
LA
MA
MI
NJ
NY
PA
TX
WA
cases relative to day −8
day
0
2
4
6
−8
−6
−4
−2
0