



Opinion

Accounting for incomplete testing in the estimation of epidemic parameters

Rebecca A Betensky* and Yang Feng

Department of Biostatistics, New York University School of Global Public Health, New York City, NY, USA

*Corresponding author. Department of Biostatistics, New York University School of Global Public Health, New York City, NY, USA. E-mail: rebecca.betensky@nyu.edu

Editorial decision 31 May 2020; Accepted 14 June 2020

Introduction

As the COVID-19 pandemic spreads across the world, it is important to understand its features and responses to public health interventions in real-time. The features of an epidemic are summarized and visualized using several measures. 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 may suffer from dependence on the underlying testing strategies within communities. That is, the numbers of cases identified are limited by the numbers of tests conducted, when cases are identified only through testing and suspected or probable cases are not identified through some other means.²

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. A decrease or flattening of the epidemic curve could be due to a decrease in the rate of infection, or it could be due to a decrease in testing, or some combination of the two. An alternative measure of the epidemic is its current doubling time,³ i.e. the time needed for prior cumulative counts of cases to double to the current count. 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.

We clarify the functional relationship between testing and the epidemic parameters, and thereby propose 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 biased. In particular, we show that the doubling time depends on the testing process through a ratio, which may be somewhat stable over short intervals of time. Importantly, this suggests that the doubling time is likely to be more robust to the unknown prevalence of testing than other measures that do not reflect testing through this ratio. We estimate this ratio for New York City under certain assumptions. Likewise, cumulative epidemic curves that are normalized to recent dates share the same desirable features.

Methods

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. We formally define the true doubling time at calendar time t to be the maximum length of time prior to t during which the cumulative number of infected cases at time t has not doubled, i.e.

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

where $C(t)$ is the cumulative number of infected cases at time t . This definition uses the fact that $C(t)$ increases over time, which means that for fixed t , $C(t)/C(t-u)$ increases as u increases, until it eventually exceeds 2, where u denotes any number of days and $t-u$ is the calendar time u days prior to time t . Due to the discreteness of measurement (i.e. daily), we cannot use an equality in this definition and find the exact time of doubling; we will handle this through linear interpolation (see below). For example, suppose that $C(10) = 95$, $C(11) = 132$, $C(12) = 148$, $C(13) = 175$, $C(14) = 181$, $C(15) = 201$, so that $C(15)/C(15-1) = 1.11$, $C(15)/C(15-2) = 1.15$, $C(15)/C(15-3) = 1.36$, $C(15)/C(15-4) = 1.52$, $C(15)/C(15-5) = 2.12$. According to the definition, $T_d(15) = 4$ because 4 is the maximum unit of time prior to $t = 15$ for which $C(15)$ is not a doubling of that prior time. Furthermore, this does not provide the exact doubling time because $C(t)/C(t - T_d(t))$ is likely <2 and not $=2$. Therefore, we interpolate using a linear approximation to estimate the exact doubling time as

$$T_d(t) + \frac{2 - C(t)/C(t - T_d(t))}{C(t)/C(t - T_d(t) - 1) - C(t)/C(t - T_d(t))}.$$

Because we do not know $C(t)$, and thus are unable to calculate $T_d(t)$, we estimate the doubling time at time t instead as

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

where $O(t)$ is the observed cumulative number of test-confirmed cases at time t . We likewise linearly interpolate this estimate to obtain an exact version. Under perfect test specificity, it follows that $O(t) \leq C(t)$. Let D_{inf} denote an individual's calendar date of infection and D_{test} denote the calendar date of testing. Then there is a simple relationship between $O(t)$ and $C(t)$, which follows from an application of Bayes theorem:

$$Prob(D_{inf} \leq t) = \frac{Prob(D_{inf} \leq t | D_{test} \leq t) \times Prob(D_{test} \leq t)}{Prob(D_{test} \leq t | D_{inf} \leq t)},$$

or equivalently,

$$Prob(D_{inf} \leq t) = \frac{Prob(D_{inf} \leq t \text{ and } D_{test} \leq t)}{Prob(D_{test} \leq t | D_{inf} \leq t)}.$$

This implies that we can estimate the true number infected by t , $C(t)$, using its conditional expected value

given the observed number infected, $E[C(t)|O(t)]$, which is given by:

$$E[C(t)|O(t)] = \frac{O(t)}{Prob(D_{test} \leq t | D_{inf} \leq t)}.$$

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:

$$\begin{aligned} T_d(t) &\cong \max \left\{ u : \frac{E[C(t)|O(t)]}{E[C(t-u)|O(t-u)]} \leq 2 \right\} \\ &= \max \left\{ u : \frac{O(t)}{O(t-u)} \times R(t, u) \leq 2 \right\}. \end{aligned}$$

where

$$R(t, u) = \frac{Prob(D_{test} \leq t-u | D_{inf} \leq t-u)}{Prob(D_{test} \leq t | D_{inf} \leq t)}.$$

$R(t, u)$ is the ratio of the probabilities of being tested among those infected by time $t-u$ relative to those infected by time 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 $R(t, u)$. 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, u) are such that $R(t, u)$ 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 u 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 individuals is constant over u units of time in $(t-u, t)$, this renders the observed doubling time estimate an accurate estimate of the true doubling time.² When $R(t, u)$ is >1 , the crude estimate of the doubling time will be an overestimate, meaning that the course of the epidemic is growing faster than it seems. This occurs when the probability of testing infected individuals is decreasing in time. When $R(t, u)$ is <1 , the crude estimate is an underestimate, and the epidemic is growing more slowly than it seems. This occurs when the probability of testing infected individuals is increasing in time; this was recognized as a possible artifact that might arise with increasing availability of diagnostic tests.²

In some cases, we may be able to estimate $R(t, u)$. Letting $D(t)$ denote the cumulative number of deaths due to the epidemic at time t , we expect the ratio, $D(t+l)/C(t)$ to be constant over time, regardless of testing strategy, for some lag time, l . This is because this quantity does not

depend on testing; the numerator includes all deaths and the denominator includes all infected cases. That is, the actual lethality of the infection should not change over time due to the underlying testing process. The ratio is known as the infection fatality rate (IFR), and is different from the case fatality rate (CFR), which includes only all symptomatic infections in the denominator. The use of a lag has been used for estimation of the CFR,^{4,5} but applies similarly to the IFR with an appropriately chosen lag. However, the ratio may still reflect the impact of the epidemic on the underlying health system, and this assumption of constant lethality will be violated if the risk of death for infected cases is increased at the height of the epidemic due to limited resources. If the assumption holds,

$$1 \cong \frac{D(t+l)}{C(t)} \bigg/ \frac{D(t-u+l)}{C(t-u)} \cong \frac{O(t)/\text{Prob}(D_{\text{rest}} \leq t | D_{\text{inf}} \leq t)}{O(t-u)/\text{Prob}(D_{\text{rest}} \leq t-u | D_{\text{inf}} \leq t-u)},$$

and thus

$$\frac{D(t+l)}{O(t)} \bigg/ \frac{D(t-u+l)}{O(t-u)} \cong R(t, u).$$

The lag time, l , acknowledges the delay from the time of measurement used in the denominator of the ratio (e.g. infection time or symptom onset time) to death.^{5,6} For the case fatality rate, in the case of COVID-19, it has been suggested that l should be 18 days, the estimated mean time from symptoms to death.⁷⁻⁹ However, recent data from the US Centers for Disease Prevention and Control¹⁰ estimate the lag from symptom onset to death to range between 12.9 and 15.3 days and the mean time from exposure to symptom onset to be 6 days, for an overall crude estimate of time from infection to death of 18.9–21.3 days. Accordingly, and given the relatively sparse data available over time, we select our lag time for the infection fatality rate to be 18 days. It is important that all deaths due to the epidemic be captured in the count $D(t)$, and not just those confirmed by testing. These data have not been reported across the USA until recently,¹¹ though they became available for New York City in mid-March,¹² and so we are able to estimate $R(t, u)$ for New York City. In that data source, deaths are labeled as ‘confirmed’ if they were preceded by a positive laboratory test for COVID-19, and ‘probable’ if there was no known laboratory test but cause of death was listed as COVID-19 on the death certificate. A limitation of this path to the estimation of $R(t, u)$ is that it can be applied only to parameters from l days prior to the current time, due to the necessity of the lag.

If the assumption of constant lethality of the epidemic does not hold due to pressures on the health care system,

for example, then the equality given above for $R(t, u)$ is actually an inequality, and we obtain a bound for $R(t, u)$, rather than a point estimate. For example, if the mortality rate has increased over the time period considered due to shortages of required medical equipment, we will have an upper bound for $R(t, u)$. Alternatively, if the mortality rate has decreased, we will have a lower bound for $R(t, u)$.

When counts of probable deaths are not available, precluding estimation of $R(t, u)$, 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 individuals 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 $u \leq (s - 1)$ and for it to remain equal to 1 for $u > s$.

An alternative, exploratory analysis is a visualization of the cumulative epidemic 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 $O(t)/O(t^*)$ vs t , for $t > t^*$. Using the same reasoning as above,

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

If $R(t, t - t^*) \cong 1$, the observed relative cumulative epidemic 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 epidemic 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 for estimation, and may or may not be accurate.

We calculated parametric bootstrap percentile 95% confidence intervals for the doubling time. We implemented this by treating each day’s new positive results as binomial, given the total tests for that day and the estimated probability of a positive test for that day. For each of 2000 bootstrap repetitions, we drew a random binomial number of new positive tests for each day. We then accumulated these over the days, to obtain a bootstrap sample of cumulative positive tests on each day. We calculated the doubling time for each repetition, and took the 2.5th and 97.5th percentiles as the bootstrap confidence interval.

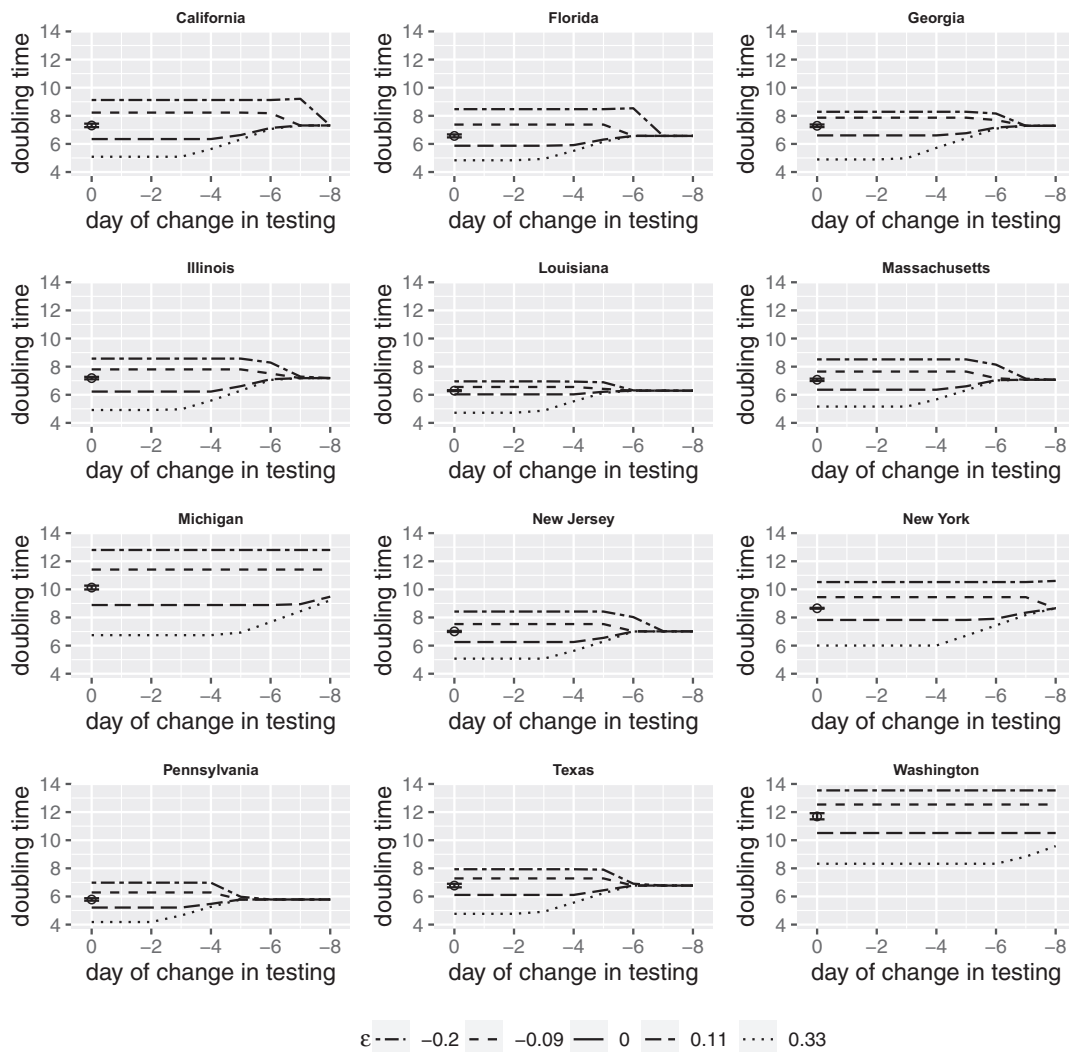


Figure 1 Doubling times with sensitivity to testing for the 12 states with the most cases as of 7 April 2020 ('day 0'). The black circle indicates the estimate under the assumption that the probability of testing infected individuals has not changed over the past 6–12 days relative to the current date of 7 April 2020 for the states depicted. We linearly interpolated the discrete-time estimates as described in the Methods. The 95% bootstrap percentile confidence interval based on the crude estimate is included. We calculated doubling times under the assumption that the probability of testing of infected individuals was constant in the past, subsequently changed on a single day in the past, and sustained that change through the current time; i.e. if c denotes that number of days prior to time t on which the change in probability of testing occurred, then $R(t, u) = 1$ for $u \leq c$ and $R(t, u) = 1 + \varepsilon$ for $u > c$. We considered values of ε of -0.2 , -0.09 , 0.11 , 0.33 , corresponding to decreases in testing of 25% (two-dashed) and 10% (dashed) and to increases of 10% (long-dashed) and 25% (dotted), respectively.

Results

Using current data on the number of positive tests in the USA and territories¹³ we have estimated the current (7 April 2020) doubling times for the 12 states with the most cases (Figure 1). The black circle indicates the estimate under the assumption that the probability of testing infected individuals has not changed over the past 6–12 days relative to the current date of 7 April 2020. We linearly interpolated the discrete-time estimates as described in the Methods. The 95% bootstrap percentile confidence interval based on the crude estimate is included. The curves depict a range of sensitivity analyses around worst-case and

best-case scenarios of over- and under-estimation of the doubling time. Specifically, we calculated doubling times under the assumption that the probability of testing of infected individuals was constant in the past, subsequently changed on a single day in the past, and sustained that change through the current time; i.e. if c denotes the number of days prior to time t on which the change in probability of testing occurred, then $R(t, u) = 1$ for $u \leq c$ and $R(t, u) = 1 + \varepsilon$ for $u > c$. We considered values of ε of -0.2 , -0.09 , 0.11 , 0.33 , corresponding to decreases in testing of 25% (two-dashed curve in Figure 1) and 10% (dashed curve) and to increases of 10% (long-dashed curve) and

25% (dotted curve), respectively. Under the model considered, if c is greater than the doubling time, $R(t, u)$ will equal 1 for the entire effective range of u , and so the estimated adjusted doubling time will equal the crude doubling time. This is seen in the figure, where the curves all converge to the crude doubling time when the day of change in testing is equal to the crude doubling time. Of note, the confidence interval for the doubling time based on the crude estimate that does not account for incomplete testing shows that the range of variability due to incomplete testing is far greater than that due to randomness. Also, the current estimates of doubling time that do not account for potential changes in testing of infected

individuals might overestimate or underestimate the true doubling time by as many as 3 days.

In Figure 2, we display the interpolated doubling times for each day, from 7 April 2020 ('day 0') back to 19 March 2020 ('day-19'), to show how they have increased over time. Again, these estimates display sensitivity to potential changes in testing of infected individuals. In this figure, we used a testing model in which we assumed that there was a single change in testing probability and it occurred on the day for which the doubling time is calculated. The solid curves depict stable probability of testing over the 20 days considered, the two-dashed curves depict a decrease of 25% in probability of testing

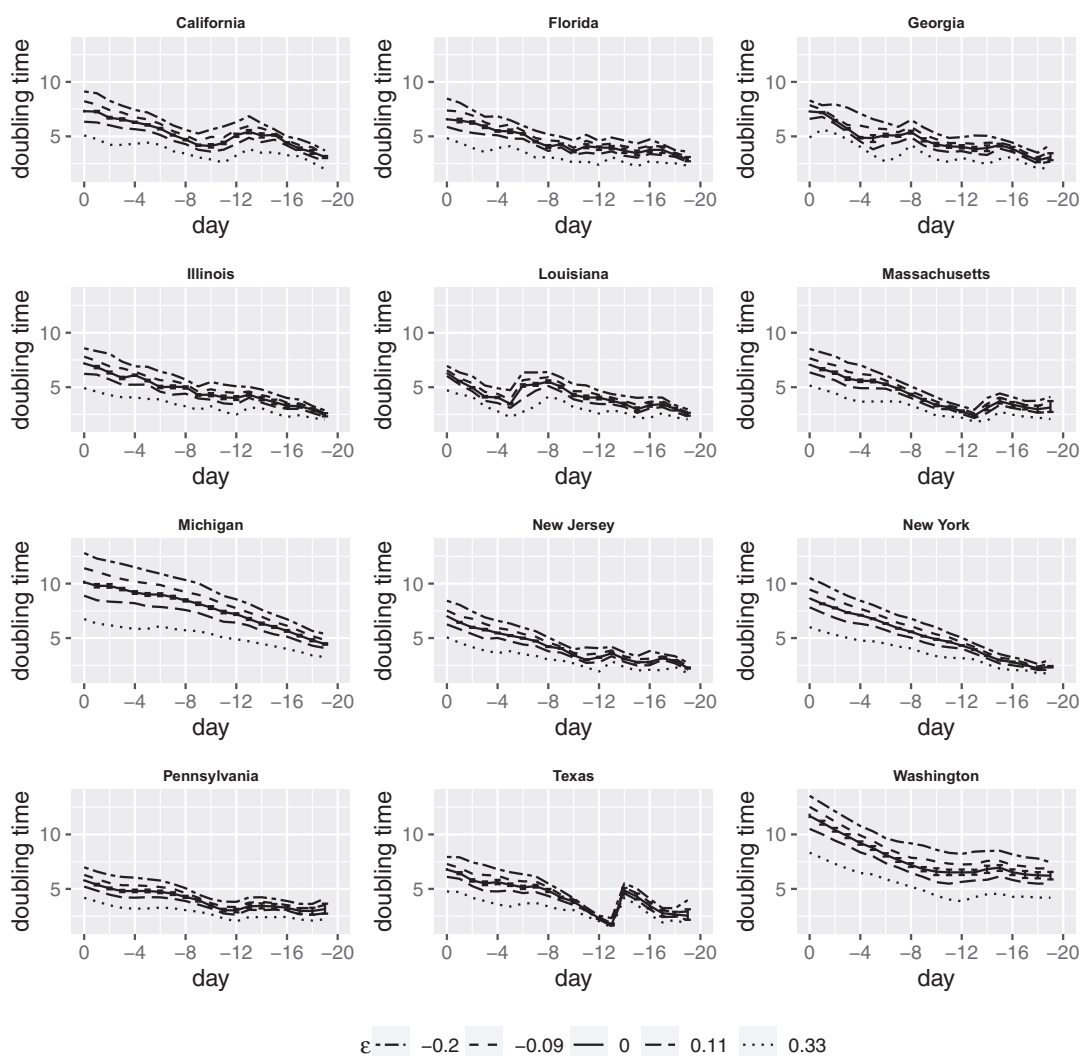


Figure 2 Doubling time on each day from 7 April 7 ('day 0') back to 19 March 2020 ('day-19'). In this figure, we used a testing model in which we assumed that there was a single change in testing probability and it occurred on the day for which the doubling time is calculated. The solid curves depict stable probability of testing over the 20 days considered, the two-dashed curves ($\epsilon = -0.2$) depict a decrease of 25% in probability of testing infected individuals on each current day versus past days, the dashed curves ($\epsilon = -0.09$) depict a decrease of 10%, the long-dashed ($\epsilon = 0.11$) curves depict an increase of 10% and the dotted curves ($\epsilon = 0.33$) depict an increase of 25%. We have included 95% bootstrap percentile confidence intervals around the solid line for the unadjusted doubling time.

individuals on each current day vs past days, the dashed curves depict a decrease of 10%, the long-dashed curves depict an increase of 10% and the dotted curves depict an increase of 25%. We have included 95% bootstrap percentile confidence intervals around the solid line for the unadjusted doubling time. The values shown at day 0 are the doubling times on 7 April 2020, and are the same values shown in Figure 1 at day of change in testing of 0. The values shown at day 4, for example, are the doubling times on 3 April 2020, and assume that testing was constant through 2 April 2020 and then changed on 3 April 2020. This provides another view, over time, of the potentially biased estimates of doubling time if testing is not considered. The very narrow confidence intervals demonstrate

that the range of variability due to incomplete testing far exceeds that due to randomness in positive tests.

In Figure 3, we display the cumulative epidemic curves from 30 March 2020 (labeled ‘day -8’) through 7 April 2020 (labeled ‘day 0’), standardized by the number of cases identified on 29 March 2020 (day 9). The solid curve assumes that testing of infected individuals has not changed in this timeframe. The dashed and dotted curves include the adjustment for differential testing and depict the scenarios in which the probability of testing of infected individuals changed on 30 March 2020 and remained constant thereafter. The changes depicted are a decrease of 25% (two-dashed curve), decrease of 10% (dashed curve), increase of 10% (long-dashed curve) and increase of 25% (dotted)

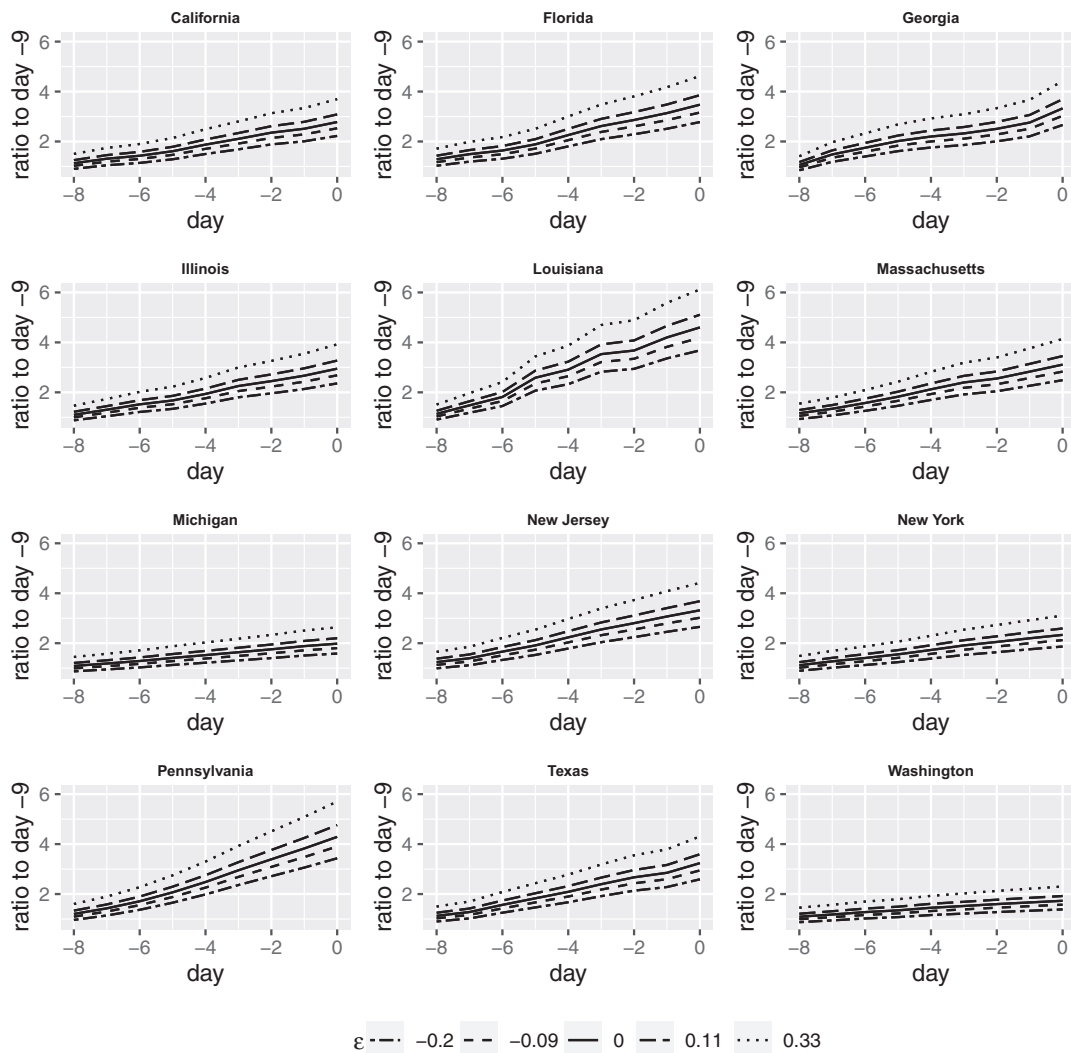


Figure 3 Cumulative epidemic curves from 30 March 2020 (day-8) through 7 April 2020 (day 0), standardized by the number of cases identified on 29 March 2020 (day-9). The solid curve assumes that testing of infected individuals has not changed in this timeframe. The dashed and dotted curves include the adjustment for differential testing and depict the scenarios in which the probability of testing of infected individuals changed on 30 March 2020 and remained constant thereafter. In particular, we considered values of ϵ of -0.2, -0.09, 0.11, 0.33, corresponding to decreases in testing of 25% (two-dashed) and 10% (dashed) and to increases of 10% (long-dashed) and 25% (dotted), respectively.

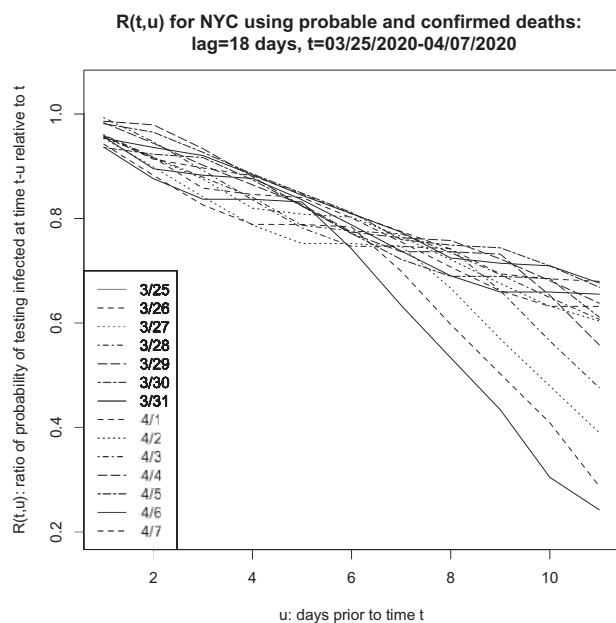


Figure 4 Estimates of $R(t, u)$ (ratio of probabilities of testing of infected individuals at time $t-u$ relative to time t) for New York City, for u ranging from 25 March 2020 through 7 April 2020 and for lag time, l , equal to 18 days. The solid curve that obtains the minimum value at $u=11$ is that for $t=25$ March 2020, and the seven curves that lie above it at $u=11$ are those for the subsequent dates, in chronological order (see legend). Beyond 1 April 2020, all curves are very close to each other.

(dotted curve). These curves demonstrate that the standardized epidemic curve is subject to bias, as a function of testing probabilities.

This series of sensitivity analyses demonstrates that a decrease in testing of infected individuals can lead to an overly optimistic view of the epidemic. In fact, this did not appear to happen in New York City in recent weeks. Figure 4 depicts $R(t, u)$ as a function of u for New York City, with each curve representing a value of t , ranging from 25 March 2020 through 7 April 2020 and for a lag time, l , of 18 days. The solid curve that obtains the minimum value at $u=11$ is that for $t=25$ March 2020, and the seven curves that lie above it at $u=11$ are those for the subsequent dates, in chronological order. Beyond 1 April 2020, all curves are very close to each other. Over this time period, $R(t, u)$, $u=1, \dots, 11$, appears to be <1 , which indicates that estimates of doubling time might be smaller, i.e. more pessimistic, than they should be. This indicates also that testing of infected individuals may have increased over time in New York City. This cannot be confirmed through examination of total numbers of tests given the unknown numbers of infected individuals. A caveat regarding these conclusions is that these estimates are imperfect due to relatively small numbers and likely incompleteness in the numbers of probable deaths. Furthermore, given the necessity of the lag of 18 days, this

approach cannot be used for real-time estimation of epidemic parameters.

The data and R code for all of our figures can be found in the [Supplementary data](#), available at *IJE* online.

Conclusion

In summary, we have illustrated formally 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. This has been addressed as a limitation in several papers on the epidemic,^{2,14} but to our knowledge has not been formally derived and investigated. In fact, in New York City, the testing of infected individuals appears to have increased slightly and so crude estimates may underestimate the slowing of the epidemic. This phenomenon was suggested as a possibility in an analysis of doubling times in provinces of China.² This may not be the case in other states or regions. For this reason, it is important to undertake sensitivity analyses around possible values of $R(t, u)$ to understand its effects on estimates of decreasing doubling times and epidemic curves, and to appropriately evaluate the effects of public health interventions. Because this is possible to do for parameters such as the doubling time and standardized cumulative epidemic curve, 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.

Supplementary data

[Supplementary data](#) are available at *IJE* online.

Funding

This work was supported by the US National Institutes of Health (R01NS094610) and the US National Science Foundation (DMS-2013789).

Author contributions

R.A.B. conceived of the study following prior analysis of doubling time by Y.F.. R.A.B. and Y.F. analyzed the data. R.A.B. and Y.F. wrote and approved the manuscript.

Data and materials availability

Data were downloaded from covidtracking.com on April 7, 2020 and from <https://github.com/nychealth/coronavirus-data> on 26 April 2020 and are available in the

[supplementary materials](#). The R code used to generate the figures is available in the [supplementary materials](#).

Conflict of interest

None declared.

References

1. Wilson EB, Burke MH. The epidemic curve. *Proc Natl Acad Sci U S A* 1942;28:361–7.
2. Muniz-Rodriguez K, Chowell G, Cheung C-H *et al*. Doubling time of the COVID-19 epidemic by province, China. *Emerg Infect Dis* 2020. doi:10.3201/eid2608.200219.
3. Nunes-Vaz R. Visualising the doubling time of COVID-19 allows comparison of the success of containment measures. *Glob Biosecur* 2020. doi:http://doi.org/10.31646/gbio.61.
4. Baud D, Qi X, Nielsen-Saines K, Musso D, Pomar L, Favre G. Real estimates of mortality following COVID-19 infection. *Lancet Infect Dis* 2020. doi:10.1016/S1473-3099(20)30195-X.
5. Dhenain M. Estimation of COVID-19 cases in France and in different countries: homogenisation based on mortality 2020. medRxiv, doi:10.1101/2020.04.07.20055913, 13 May 2020, preprint: not peer reviewed.
6. Flaxman S, Mishra S, Gandy A, *et al*. Estimating the number of infections and the impact of nonpharmaceutical interventions on COVID-19 in 11 European countries. *Imperial College London*. 2020. doi:10.25561/77731.
7. Ruan Q, Yang K, Wang W, Jiang L, Song J. Clinical predictors of mortality due to COVID 19 based on an analysis of data of 150 patients from Wuhan, China. *Intensive Care Med* 2020;46: 846–3.
8. Verity R, Okell LC, Dorigatti I *et al*. Estimates of the 2 severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect Dis* 2020;S1473-3099:30243–7.
9. Zhou F, Yu T, Du R *et al*. Clinical course and risk factors for mortality of adult in patients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet* 2020;395:1054–62.
10. Centers for Disease Control and Prevention. COVID-19 Pandemic Planning Scenarios. 2020. <https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html#table-2> (28 May 2020, date last accessed).
11. Centers for Disease Control and Prevention. Cases in the U.S. 2020. <https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html> (28 May 2020, date last accessed).
12. Github. NYC Coronavirus Disease. 2019 (COVID-19) Data. 2020. <https://github.com/nychealth/coronavirus-data> (28 May 2020, date last accessed).
13. The COVID Tracking Project. Data API. 2020. <https://covidtracking.com/api> (28 May 2020, date last accessed).
14. Wu JT, Leung K, Bushman M *et al*. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nat Med* 2020;26:506–10.